A Genomic and Proteomic Investigation Into the *Clostridium botulinum* Neurotoxin Complex

Thesis

A Genomic and Proteomic Investigation into the *Clostridium botulinum* Neurotoxin Complex

Philip Matthew Ashton BSc

Affiliated Research Centre – Health Protection Agency

This work is offered in part fulfilment of a PhD in Biological Sciences in November 2013

DATE of SUBMISSION: 26 NOVEMBER 2013

DATE of AWARD: 27 MAY 2014

Acknolwedgements

I would like to thank all the people whose forebearance and understanding was so heavily relied upon during the writing of this thesis. Specificially Nikki Rockell for the first 4 years and Cat Anscombe for the last 6 months, which are obviously the most important.

I would also like to thank:

- For proteomic assistance, Min Fang, Tom Gaulton, Raju Misra and Russell Jones.
- For microbiological assistance, Ijay Nwafor, Charles Ohai and David Wooldridge.
- For bioinformatic assistance, Raju Misra, Anthony Underwood, Richard Myers and Tim Dallman.

I could not have finished (or started!) this PhD without the support, advice, and guidance of my three excellent supervisors Kathie Grant, Mike Peck and Saheer Gharbia.

I would also like to thank the people who have helped me get where I am today over the longer term. Specifically, my father Richard and brother James.

Contents

List of figures

Abstract

Clostridium botulinum and some strains of *C. baratii* and *C. butyricum* produce one of the most potent toxins known to man, botulinum neurotoxin, and are responsible for the disease botulism. This severe neuroparalytic disease is the result of botulinum neurotoxin negotiating a complex path to the cholinergic nerve endings. There, it interferes with the release of excitatory neurotransmitters resulting in flaccid paralysis and if untreated, death. The neurotoxin, itself a multi-faceted protein, does not act alone but is produced as part of a large, hetero-multimeric complex with the associated non-toxic proteins. This complex, known to protect the toxin from acids and proteases in the gut, has recently been suggested to play a more active role in toxicity. Here, the proteins from the lesser-studied toxin complex type (the OrfX type) are shown for the first time to share sequence similarity and synteny with clusters of proteins that are co-localised with various putative toxin genes in diverse other species. The extracellular supernatant proteome of *C. botulinum* is characterised and mined for potential novel virulence factors, with metabolic cost of extracellular protein being highlighted as a potential marker of virulence associated extracellular proteins. The supernatant of 22 clinical strains of *C. botulinum* were investigated for the presence of the toxin complex; all toxin complex components were identified in the majority of strains indicating the importance of these proteins in the causation of botulism. A relationship between OrfX-encoding strains and infant botulism was also uncovered in clinical strains from the UK. Hypotheses to explain this association are explored. Finally, the transcriptome of *C. botulinum* was investigated using RNA-sequencing. This uncovered a complex and diverse picture of transcription in *C. botulinum* and raised questions as to the role of the alternative sigma factor BotR in the regulation of *bont*.

1. Introduction

1.1. Clostridium – The Genus

The genus *Clostridium* has traditionally been classified as anaerobic or aerotolerant, endospore-forming rods that are gram positive (at least in the early stages of growth) (Collins et al, 1994). As a result of these broad defining criteria, *Clostridium* is a large and diverse group with a wide range of G+C contents (22-55%) although the toxigenic species have a much narrower GC range (24-29%) (Hatheway, 1990). Designation of a bacterial isolate as a *Clostridium* is dependent on it fulfilling 4 criteria (1) the ability to produce endospores (2) having an anaerobic energy metabolism not involving oxygen as an electron acceptor (3) the inability to reduce sulphate to sulphide – this characteristic separates them from *Desulfotomaculum* (4) Gram positive cell wall, at least in the early stages of growth.

Bergey's Manual of Systematic Bacteriology; The Firmicutes (Vos et al, 2009), lists 200 species in the genus *Clostridium*. When the phylogeny of the genus is analysed using gene or protein sequences the 200 constituent species do not form a monophyletic group (Collins et al., 1994; Gupta and Gao, 2009). In a seminal paper on the phylogeny of *Clostridium* by Collins et al (1994) the genus was divided into 19 clusters based on 16S rDNA sequences. Three of these *Clostridium* clusters account for 70% of the clostridia present in the normal faecal flora of humans (Finegold et al, 2002). Included in cluster I were *Clostridium butyricum* (the genus type species) and the medically important species *C. botulinum, C. tetani, C. perfringens* and *C. baratii*. Collins et al. suggested that the definition of the genus *Clostridium* be restricted to species in cluster I only, this would reduce the number of species in the *Clostridium* genus by over 50%. In

addition to the 16S rDNA based study of Collins et al., the phylogeny of *Clostridium* was investigated by concatenating 37 conserved proteins and analysing them for the presence of indels (Gupta and Gao, 2009). In the phylogenies generated by this methodology, the cluster I *Clostridium* species were clearly distinguishable from other clusters, replicating the results obtained in 16S rDNA phylogeny.

In addition to the many pathogens in the genus *Clostridium* there are also some industrially important species such as *C. acetobutylicum* (belongs to cluster I) that is capable of fermenting starch to acetone, butanol and ethanol.

1.2. *Clostridium* – The Pathogens

Clostridia are among the most widely distributed pathogens in the environment. There are at least 35 pathogenic *Clostridium* that are found in a range of habitats including soil and the gut microflora of humans and other animals (Peck, 2009). *Clostridium* produce more types of protein toxins than any other genus of microorganisms with at least 20 identified toxins produced by 15 different species. These toxins include neurotoxins, proteases, cytotoxins, lipases, collagenases, necrotizing toxins, ADP-ribosyltransferases and neuraminidases (Hatheway, 1990).

Clostridium perfringens is the most frequently identified *Clostridium* pathogen, accounting for 20% of *Clostridium* encountered in clinical specimens (Allen et al, 2003). It is implicated in a wide variety of clinical conditions including but not limited to myonecrosis, cellulitis; intra-abdominal sepsis, septicaemia, intra-vascular hemolysis and is one of the most common bacterial causes of food

poisoning gastroenteritis in the USA and elsewhere in the developed world (Thomas et al., 2013; Gormley et al., 2011; Olsen, 2000). *C. perfringens* food poisoning is associated with food products that have been time and temperature abused. Clostridial myonecrosis, otherwise known as gas gangrene, is a life-threatening, rapidly developing condition caused by, primarily, *C. perfringens* although other *Clostridium* species can be responsible.

Although tetanus is a relatively minor problem in the developed world it is still a major problem in the developing world. The World Health Organisation estimate that 59000 newborn children died from neonatal tetanus in 2008 (http://www.who.int/immunization_monitoring/diseases/MNTE_initiative/en/index.html). *Clostridium tetani* spores are widespread in the environment and can infect anaerobic wounds (Popoff, 1995).

Clostridium difficile is the major cause of antibiotic associated diarrhoea and pseudo-membranous colitis as well as being the most frequent nosocomial diarrhoeal pathogen (Lessa et al., 2012). This rise has been attributed to the emergence of a previously rare, hypervirulent strain, *C. difficile* O27 that shows increased resistance to fluoroquinolones and increased toxin production (Lessa et al., 2012).

Botulism is caused by *Clostridium botulinum* and rare strains of *C. baratii* and *C. butyricum* which produce the most lethal protein toxin known to man, botulinum neurotoxin (BoNT).

1.3. Botulism and *Clostridium botulinum*

The disease botulism was first described in 1821 (Erbguth, 2004) but it wasn't until Emile van Ermengem investigated a cluster of food-borne botulism at a funeral in Belgium in the 1890s that *Clostridium botulinum* (then known as *Bacillus botulinus*) was established as the causative agent of botulism (1979 translation of van Ermengem, 1897). It was not until the latter half of the 20th Century that the ability of *C. botulinum* to cause wound (Davis et al., 1951) and infant botulism (Pickett et al, 1976) was established.

Analysis of 16S rDNA sequences has shown that *C. botulinum* actually consists of four distinct lineages, separated by a large number of other clostridial species (Collins et al., 1994). This finding is additionally supported by DNA hybridisation data and Amplified Fragment Length Polymorphism (AFLP) analysis (Collins and East, 1998; Hill et al, 2007). Each of the four lineages has other, non-*C. botulinum* species that are much more closely related to it than the other *C.* botulinum lineages (Table 1). However, historical precedent combined with the clinical relevance of the botulinum toxin producing phenotype have resulted in a single species name being applied to what, in phylogenetic terms, are four separate species. Additionally, the botulinum toxin has also been identified in the genomic background of other clostridial species e.g. *C. butyricum* and *C. baratii* thereby rendering the definition of *C. botulinum* as all strains that produce BoNT not factually accurate.

Eight serologically distinct botulinum neurotoxin serotypes (types A-H) have been described with six of eight serotypes having subtypes (Montal, 2010; Barash & Arnon, 2013). These strains of *C. botulinum* may produce one or two different

serotypes – strains tend to produce predominantly one type and less of the other, this is expressed as e.g. Ab. Strains that encode the genes for two toxin types but only produce one are expressed as e.g. A(B). Four of the BoNT serotypes are associated with human disease; types A, B, E and F, while C and D are associated with animal disease and type G has not been definitively associated with any human or animal disease (Arnon et al., 2001). Type H was very recently identified (Barash & Arnon, 2013). The workers who identified it have not yet released the amino acid or nucleotide sequence of the toxin as they fear that a terrorist group could synthesise the toxin before appropriate anti-toxin was developed (Dover et al., 2013).

The four lineages (I-IV) of *C. botulinum,* in addition to forming distinct genetic clusters also form four physiologically distinct groups (Table 1) (Allen, 2003). The physiology of these lineages differs in toxin type production, growth temperature range, proteolytic activity, the host range in which they cause disease and the sensitivity of their spores to heat treatment (McLauchlin & Grant, 2007). Thus, in addition to the genetic differences highlighted, the phenotypic differences between these groups justify their reclassification as separate species. Only group I and II *C. botulinum* are commonly associated with human disease. One characteristic distinguishing group II organisms is the ability to grow and produce toxin at refrigeration temperatures. This heightens the risk of food contamination in the modern food supply chain where refrigeration is frequently relied upon to prevent bacterial growth (Chen et al, 2008). Group III strains cause disease in animals and birds and can grow at temperatures of 15-40°C. There has only been one reported case of human botulism caused by a group III organism, an infant botulism case in Japan (Oguma et al, 1990). Group IV *C. botulinum* is also known as *Clostridium*

argentinense (Suen et al, 1988). Group IV organisms have not been extensively studied as they are rare and have caused no confirmed cases of human botulism. However, they have been inconclusively linked with Sudden Infant Death Syndrome (SIDS) (Sonnabend et al, 1985).

Table 1: Characteristics of *C. botulinum* groups I-IV

	Group			
	I	II	III	IV
Neurotoxin types	A, B, F	B, E, F	C, D	G
Human disease	Yes	Yes	No	No
Growth temperatures				
Minimum °C	10	3.3	15	ND
Optimum °C	35-40	18-25	40	37
Minimum pH for growth	4.6	5	5	ND
Proteolytic activity	Yes	No	No	Yes (weak)
Related non-neurotoxigenic *clostridia*	*C. sporogenes*	*C. beijerinkii*	*C. novyi*	C. histolyticum
Location of neurotoxin genes	Chromosome or plasmid	Chromosomal	Phage	Plasmid
Overall G+C content (mol %)	26-29	27-29	26-28	28-30
$D_{100°C}$* of spores (min)	25	<0.1		
$D_{121°C}$* of spores (min)	0.1-0.2	<0.001		

ND = not determined

*D (decimal reduction) values are the time taken to kill 90% of organisms at a specific temperature.

1.4. Botulism Epidemiology

Botulism is a rare but serious disease that can lead to death. There are three main forms of human botulism; food borne botulism, wound botulism and infant botulism. In addition to these main forms there are other, exceedingly rare forms including inhalational botulism, adult intestinal botulism and iatrogenic botulism.

1.4.1. Food borne botulism

Food-borne botulism occurs when food containing preformed botulinum neurotoxin is consumed. Botulinum neurotoxin is typically generated when *C. botulinum* spores contaminate the food and are able to germinate and grow due to a suitable anaerobic environment and produce toxin. Even small outbreaks of botulism constitute public health emergencies due to the extreme toxicity of botulinum neurotoxin and the potential for wide distribution of contaminated food possibly resulting in large numbers of victims. Cases of botulism can result in significant strain on public health and acute care provision as well as having significant medical and financial impact. It has been estimated that the average financial impact of a case of botulism is $30 million (Peck, 2009). There were 6 reported episodes of food-borne botulism in the UK between 1989-2005 resulting in 33 reported cases and 3 deaths (McLauchlin et al., 2006) with a total of 62 cases being recognised between 1922 and 2005. Since 2005 there have been three incidents of food-borne botulism, resulting in 5 cases. The largest of these incidents was a familial cluster of botulism cases associated with a commercially prepared Korma sauce (Browning et al., 2011), while artisanally produced olives were implicated in another case (K. Grant, Personal Communication).

The largest recorded outbreak of botulism in the UK was in 1989 when 27 cases were associated with commercially prepared hazelnut yoghurt. Yoghurt would not normally be considered a botulism risk food due to it's high acidity, however, it is thought that BoNT was produced by *C. botulinum* in the hazelnut conserve used to flavour the yoghurt. Other foods associated with botulism in the UK have included home preserved mushrooms in oil from Italy, commercially prepared hummus stored at room temperature for several weeks before consumption, various traditionally prepared foods from Georgia and home preserved pork from Poland (McLauchlin et al., 2006). The deadliest outbreak of food-borne botulism in the UK was also the first reported outbreak. Consumption of contaminated duck paste at Loch Maree in Scotland caused 8 cases of botulism that all resulted in death (Leighton, 1923).

One reason for the low incidence of food-borne botulism is the adoption of a 'botulinum cook' as a standard part of the production of low-acid canned goods. This involves heating the canned food to 121°C for 3 minutes. However, changes in eating habits such as higher consumption of minimally processed foods stored at refrigeration temperatures has changed the risk profile of foods that could be potentially associated with botulism (Peck, 2009).

Certain European countries have much higher numbers of food-borne botulism cases than the UK. Poland has one of the highest rates of food-borne botulism in the world with nearly 1301 incidents between 1984-87 resulting in 1791 cases (Table 2). This extremely high number is due to a culture of home food preservation, social and economic turmoil resulting in food shortages in the 1980s

and meticulous recording in 1984-87 (Hauschild, 1993). Between 1988 and 1998 Italy had 412 cases of foodborne botulism, Germany 177, Spain 92 (Galazka & Przybylska 1999) while the USA had 597 cases between 1971-89 (Hauschild, 1993).

More recent data (1990-2000) provides a detailed breakdown of the foods implicated in food-borne botulism in the USA (McLauchlin and Grant, 2007). The most frequently implicated source was non-commercial preserved fish or marine mammals which resulted in 92 cases, these are typically associated with indigenous populations in Alaska. Home canned vegetables caused 70 cases with asparagus and olives among the produce implicated (McLauchlin and Grant, 2007). An outbreak in 2006 involved temperature abused, commercial carrot juice that affected 6 people, all of whom required mechanical ventilation (Sheth et al., 2008). Only nine cases of botulism were associated with home preserved meat (McLauchlin and Grant, 2007). There is a close association between the frequency/toxin type of a botulism outbreak (incidents involving one or more botulinum cases) and the occurrence and toxin type of *C. botulinum* spores in that particular environment (Hauschild, 1989).

Different botulinum toxin types are associated with different mortality rates in all forms of botulism, botulinum toxin type A has a significantly higher mortality in than type B (Hauschild, 1993). This is reflected in food borne botulism mortality rates associated with different toxin types. In Poland 94% of botulism cases are caused by toxin type B and the mortality rate is 3%. By contrast, in Argentina 77% of cases are caused by toxin type A and there is a mortality rate of 36%. This pattern

is repeated in other countries, France has 97% type B cases and a mortality rate of 2% while China has 93% type A cases and a 13% mortality rate (Table 2A, 2B), although it is difficicult to control for the effect of different standards of healthcare in the two countries (Hauschild, 1993). The difference in mortality rate between Argentina and China may be due to the relatively small number of cases in Argentina (36) compared with China (4377). There also appears to be a relationship between food type and toxin type (Table 2C). In countries that have a majority of botulism caused by type B strains, the majority of cases are associated with meat products while countries with a higher incidence of type A botulism have more cases associated with fruit and vegetables. Type E cases are strongly associated with the consumption of fish and type E producing strains are typically found in aquatic and marine environments (Hauschild, 1993).

Table 2: Foodborne botulism epidemiology. Table adapted from Hauschild et al., 1993. Dates in (B) and (C) same as (A).

(A) Summary of the number of outbreaks, cases and fatalities associated with botulism.

	Period	Outbreaks	Cases	Fatalities (%)	Outbreaks/yr	Cases/yr
United States	1971-89	272	597	63 (11)	14	31
Canada	1971-89	79	202	28 (14)	4	11
Argentina	1980-89	16	36	13 (36)	2	4
Poland	1984-87	1301	1791	46 (3)	325	448
Czechoslovakia	1979-84	17	30	0	3	3
Hungary	1985-89	31	57	1 (2)	6	11
Yugoslavia	1984-89	12	51		2	8
Belgium	1982-89	11	25	1 (4)	1	3
France	1978-89	175	304	7 (2)	15	25
Spain	1969-88	63	198	11 (6)	3	10
Portugal	197-89	24	80	0	1	4
Norway	1961-90	19	42	3 (7)	<1	1
U.S.S.R	1958-64	95	328	95 (29)	14	47
China	1958-83	986	4377	548 (13)	38	168
Japan	1951-87	97	479	110 (23)	3	13

(B) The percentage of cases associated with different toxin types

	Outbreaks with type identified	A (%)	B (%)	E (%)	Other (%)
United States	252	61	21	17	0.4 (F)
Canada	76	4	8	88	0
Argentina	13	77	8	0	15 (Af)
Poland	830	3	94	3	0
Czechoslovakia	6	17	83	0	0
Hungary	31	0	100	0	0
Belgium	11	0	55	9	36 (Bc)
France	171	0	97	2	0.6 (Ab)
Spain	36	0	92	3	6 (Ab)
Portugal	18	0	100	0	0
Norway	19	0	47	47	5 (F)
U.S.S.R	45	33	38	29	0
China	733	93	5	1	0.4 (Ab)
Japan	97	2	2	96	0

(C) Type of food associated with outbreaks.

Food type or source (%)	Outbreaks	Meat	Fish	Fruit and veg	Mixed	Home preserved	Commercial
United States	222	16	17	59	9	92	8
Canada	75	72	20	8	0	96	4
Argentina	14	29	21	36	14	79	21
Poland	1500	83	12	5	0	75	24
Czechoslovakia	14	72	7	14	7	100	0
Hungary	28	89	0	4	7	100	0
Yugoslavia	8	100	0	0	0	100	0
France	123	89	3	6	2	88	12
Spain	48	38	8	60	0	90	10
Portugal	23	91	9	0	0	100	0
Norway	19	16	84	0	0	100	0
U.S.S.R	83	17	67	16	0	97	3
China	958	10	0	86	4	N/A	N/A
Japan	95	0	99	1	0	98	2

1.4.2. Wound botulism

Wound botulism occurs when a wound, typically caused by either traumatic injury or puncture associated with injected drug use (IDUs), becomes contaminated with spores of *Clostridium botulinum.* If the spores are present within an anaerobic wound, they can germinate, grow and release botulinum toxin *in vivo*. The first case of wound infection caused by *C. botulinum* was reported in 1951 in a trauma wound patient from California (Davis et al, 1951). Infection of trauma wounds or post-operative infection was the only recognised cause of wound botulism until 1982 when an injecting drug user who was subcutaneously injecting cocaine ('skin popping') was diagnosed with wound botulism in New York (Weber et al, 1993). This led to the recognition of a new type of wound botulism associated with drug use – the majority of which were due to injecting drug users although there have been rare cases that are thought to be associated with sinusitis/intranasal cocaine abuse. The recognition of this new cause of wound botulism was due to a dramatic increase in the number of cases of wound botulism being diagnosed in California (Figure 1). The first reported case of wound botulism in an injecting drug user in the UK was in 2000, nearly 20 years after the phenomenon was recognised, although more cases quickly accrued afterwards. There were 33 reported cases in the UK between 2000-2002 (Brett et al, 2004). Wound botulism was the most commonly reported form of botulism in the UK between 2000-2009 with 162 cases (Figure 2). However, since then the number of wound botulism cases has declined and now there are approximately 5 cases of wound botulism per year. Whether there was unreported wound botulism prior to 2000 is unclear. Increased recognition by clinicians is likely to have played a role in the increase of detection rate with outbreaks also be due to contamination of specific batches of heroin (Akbulut et al, 2005).

C. argentinense was isolated from a trauma patient with a compound fracture who had developed wound botulism in North Carolina, however the patient responded well to a mix of type A/type B antitoxin which makes it unlikely that the causative agent of the wound botulism was *C. argentinense*. This highlights the complexity of diagnosing disease caused by a spore forming organism like *C. botulinum* (Taylor et al, 2010).

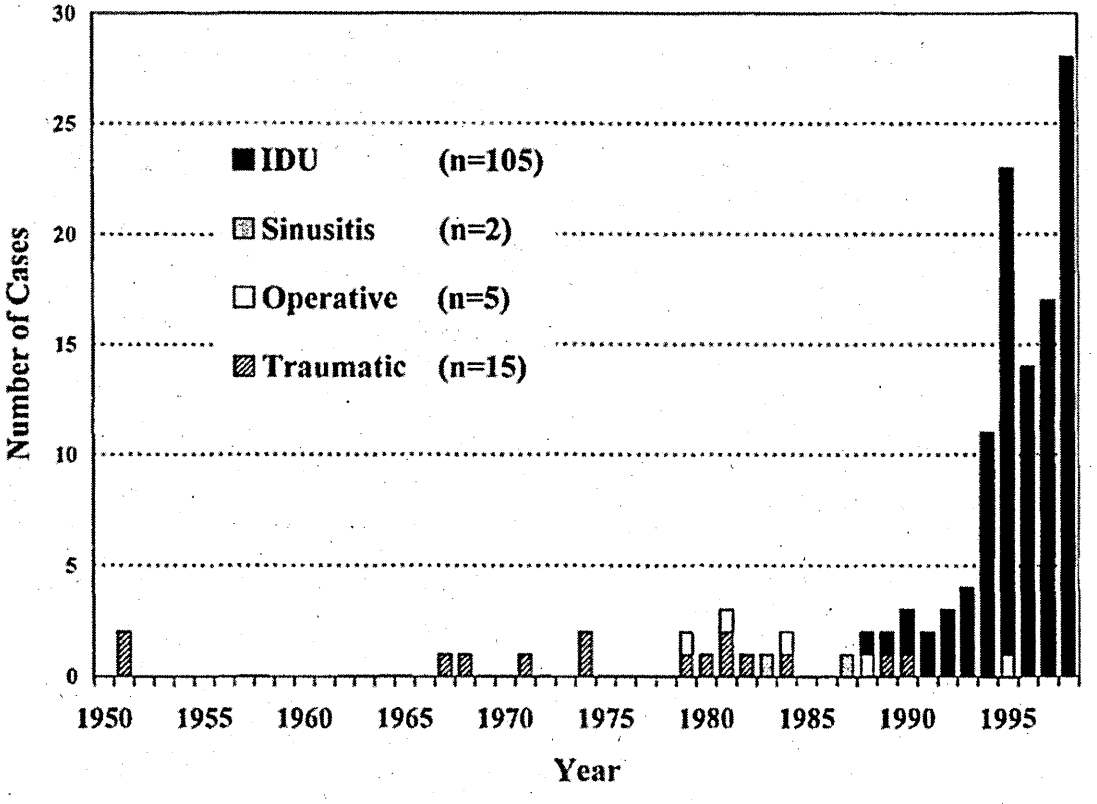


Figure 1: Incidence of wound botulism in California by wound type (IDU = Injecting Drug User), adapted from Werner et al, 2000

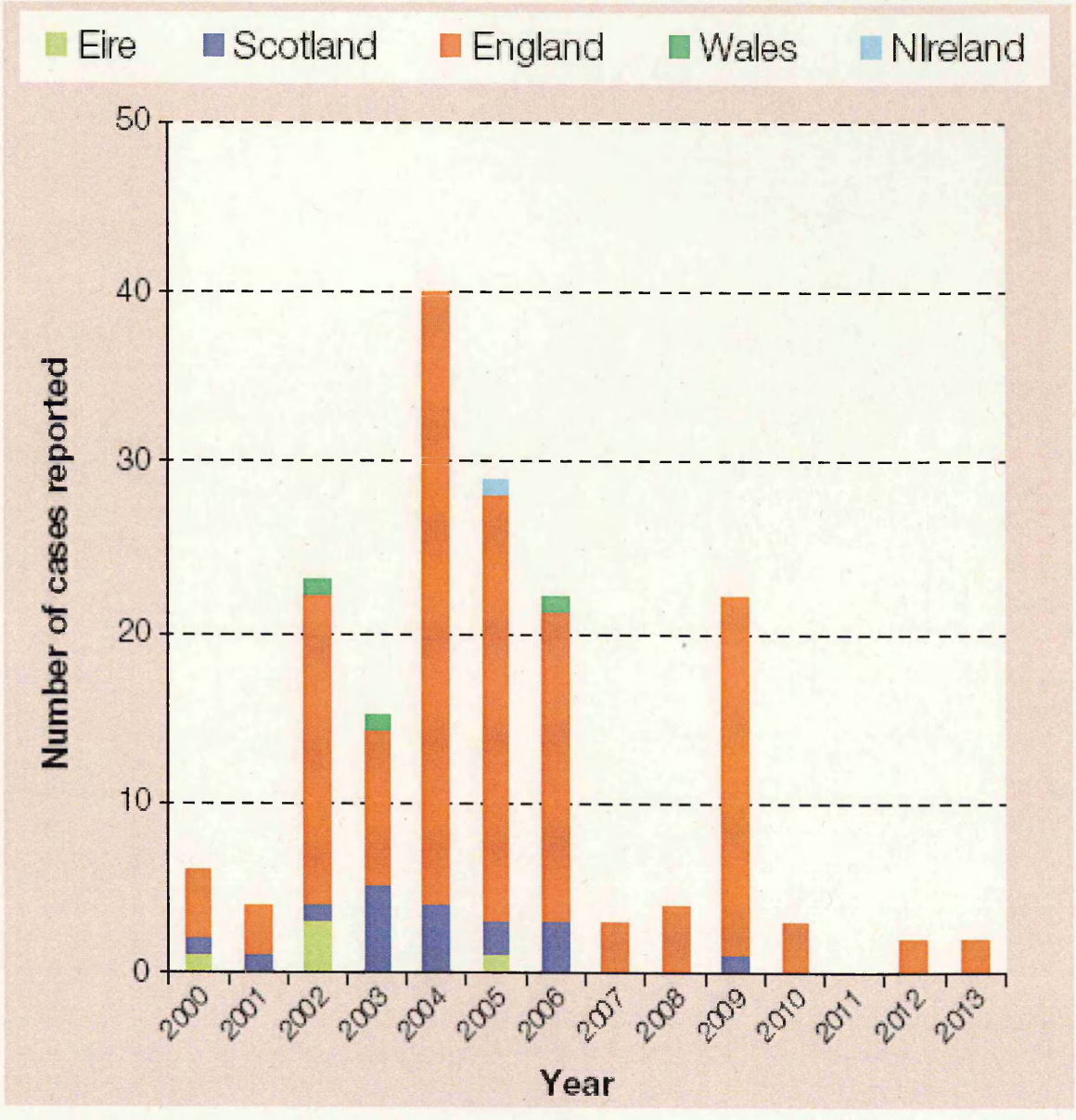


Figure 2: Incidence of wound botulism in the UK and Eire 2000-13, personal communication K. Grant and K. Cullen.

1.4.3. Infant botulism

Infant botulism is a toxicoinfection caused by the colonisation and toxin production in the gut of infants less than 12 months old by *C. botulinum*. It was first recognised as a distinct clinical entity by Midura and Arnon in 1976. To date, infant botulism has been reported in 26 countries with the USA, Argentina, Australia, Italy, Canada and Japan reporting the most cases in descending order (see Table 3).

The majority of infant botulism cases are caused by *C. botulinum* groups I and II producing toxin types A, B, E, Ab, Ba, Bf (the designation Ab means the strain predominately produces type A but also produces small amounts of type B) and in one case in Japan a group III organism producing type C (Koepke et al, 2008). Toxigenic, non-*C. botulinum* species including *C. baratii* producing BoNT/F and *C. butyricum* producing BoNT/E have also been associated with infant botulism in the USA, Japan, Italy and the UK (Fenicia and Anniballi, 2009; Personal Communication, Grant). *C. argentinense* producing type G toxin has been associated, inconclusively, with a sudden and unexplained infant death in Switzerland (Sonnabend et al, 1981). Most infant botulism is caused by group I organisms, the fact that their optimal growth temperature is 35-40 °C while group II strains optimal growth temperature is 18-25 °C would place the group II organisms at a fitness disadvantage to colonise the infant gut.

More than 82% of the reported worldwide cases of infant botulism have occurred in the United States (Koepke et al., 2008). The USA had 2419 cases reported between 1976 and 2006 (2.1 cases per 100 000 live births), with roughly equal

numbers of males and females affected (Koepke et al., 2008). In the USA the mean age of onset was 13.8 weeks with 90% of cases falling between 2 weeks and 6 months although there was a range from less than a week old to 2 years of age (Koepke et al., 2008).

The fact that most cases of infant botulism occur in infants less than one year of age reflects the ability of the organism to colonise the gut at this stage in the infant's life. The susceptibility of the infant gut to colonisation is a result of an immature gut flora combined with an event that perturbs the gut flora, e.g. weaning, change of milk formula or antibiotic treatment (Grant et al, 2009; Dodds, 1992b). Infant botulism must be the result of a complex host-pathogen interaction as older infants, children and adults are frequently exposed to *C. botulinum* spores in the environment and foods without developing disease. Additionally, infant botulism remains a rare disease despite the ubiquitous presence of *C. botulinum* in the environment, therefore risk of disease development is likely a factor of exposure and host susceptibility caused by perturbation of immature infant gut. An additional environmental factor influencing the exposure of infants to *C. botulinum* is a dry, dusty environment. The presence of building works, farms or other activities that may create large amounts of dust can constitute risk factors (Long et al, 1985). There was double the number of usual cases in the month following the Northridge earthquake in California in January 1994, possibly due to more spores being released into the air (Underwood et al, 2007). Another piece of evidence in favour of the environment being the source of spores that lead to infant botulism is the association between the dominant spore type in an area and the major causative type of infant botulism cases in that area. Type A cases predominate west of the Rocky Mountains in the US while type B cases predominate in the east, matching

the environmental distribution of *C. botulinum* spores in these areas (Fenicia and Anniballi, 2009; Dodds, 1992b).

Honey is the only food proven, both epidemiologically and microbiologically, to be associated with infant botulism. Warning labels on pots of honey stating that honey should not be given to children less than 1 year old were introduced in the USA in the 1970s. The number of infant botulism cases associated with honey there has fallen from 39.7% in the 1970s to 4.7% in the 2000s (Koepke et al, 2008). Similar warning labels on honey pots are now in place in many other countries including the UK, Japan, Canada and Australia.

Table 3: Reported cases of infant botulism according to country, 1976-2006, adapted from Koepke et al., 2008

Location	Time period	Total No.	Type A	Type B	Other types	Toxin type not reported
Global total, excluding United States	1976-2006	542	437	56	12	19
Asia						
China	1986-1989	2	0	1	0	1
Japan	1986-2006	22	14	3	2	3
Taiwan	1987	1	0	1	0	0
Australia	1978-2006	32	12	15	1	4
Europe						
Czech Republic	1979	1	0	1	0	0
Denmark	1995-2000	2	0	0	1	1
France	1983-2006	4	1	3	0	0
Germany	1993-200	4	2	0	0	2
Greece	2006	1	1	0	0	0
Hungary	1995-2002	2	0	0	1	1
Italy	1984-2006	24	4	17	5	0
Netherlands	2000-2005	3	1	2	0	0
Norway	1997-1999	4	4	0	0	0
Spain	1985-2002	9	2	2	0	5
Sweden	1985-2006	3	2	0	1	0
United Kingdom	1978-2001	5	2	2	1	0
Middle East						
Israel	1994-2006	2	0	2	0	0
Kuwait	2005	1	0	0	0	1
Yemen	1989	1	0	1	0	0
Canada	1979-2006	27	22	5	0	0
Mexico	2001	1	1	0	0	0
United States (exc Rhode Island)	**1976-2006**	**2419**	**1079**	**1310**	**28**	**2**
South America						
Argentina	1982-2005	366	366	0	0	0
Chile	1984-1995	3	2	0	0	1
Venezuela	2000	1	0	1	0	0

Infant botulism is rare in the UK with only 16 cases reported since 1978, 10 of these have occurred since 2006 with two unrelated cases in a single month in 2007. In terms of risk factors, 5 infants were being weaned, 7 had received honey and one had received antibiotic treatment (Grant et al., 2009; K. Grant, Personal Communication).

A link between infant botulism and Sudden Infant Death Syndrome (SIDS) was first suggested in 1976 by Midura and Arnon. Studies of necropsy specimens from SIDS cases have found varying results depending on the location of the study. SIDS cases were screen for *C. botulinum* toxin or organism in a study in Australia that included 248 infants. None of the infants were positive for toxin or organsim (Byard et al, 1992). Other studies have found more significant results - 4.3% of 280 patients in a study in California (Arnon et al, 1978) and 15% of 70 cases in a study in Switzerland (Sonnabend et al, 1985) were positive for either botulinum toxin or organism. The Swiss study that detected toxin in 15% of cases was unusual in that the toxin types C and G were also screened for, this study was also the first report of type G being associated with human disease (Sonnabend et al, 1985; Sonnabend et al, 1981). This association is however, supported by limited evidence and remains controversial. There are likely many different causes of SIDS, of which infant botulism appears to be one in some areas of the globe.

1.4.4. Other forms of botulism

There are four naturally occurring forms of botulism, three of which (food, wound and infant) are covered above. The fourth, adult intestinal colonisation is very rare (Fenicia & Anniballi, 2008), and has been associated with host factors such as Crohn's disease or having undergone bowel surgery which disrupt the intestinal

flora, allowing *C. botulinum* spores the chance to germinate and produce toxin within the host (Sheppard et al., 2012).

Two documented man-made forms of botulism are inhalational botulism and iatrogenic botulism (Sobel, 2005). There has been one known incident of inhalational botulism, in German laboratory workers in 1962 (Sobel, 2005). Inhalational botulism is also a possible route of transmission for deliberate release of botulinum toxin (Arnon et al., 2001). Iatrogenic botulism is caused by injection of toxin for cosmetic or therapeutic purposes. Recommended doses for cosmetic treatment are too low to cause systemic disease but injection of unlicensed, highly concentration botulinum toxin has caused severe botulism (Sobel, 2005). There have been anecdotal reports of botulism after injection of botulinum toxin to control muscle movement disorders as the doses used for this are much higher than the recommended cosmetic dosage (Sobel, 2005).

Botulism also affects many animal species. Group III strains of *C. botulinum* producing toxins C and D cause botulism in livestock animals and birds. Type C toxin affects most animal species, notably wild ducks, pheasant, chickens, mink, horses and cattle (Degernes, 2008; Wylie & Proudman, 2009). Type D is best known for outbreaks in cattle with the usual source of intoxication is spoiled silage or fodder (Lindstom et al., 2010). Prevalence in animals is not well known but estimates of 10 000 to 50 000 birds a year die from botulism. Dogs, cats and pigs are comparatively resistant to botulism for reasons that remain unclear (Critchley, 1991).

1.4.5. Symptoms, diagnosis and treatment of botulism

Food-borne botulism has a period between toxin consumption and symptom development that ranges from 6 hours to 10 days, with symptoms generally apparent 18-36 h after consumption of toxin (McLaughlin & Grant, 2007). Typically, the first symptoms will be gastrointestinal (e.g. vomiting and diarrhea) followed by other symptoms such as blurred/double vision, drooping eyelids, difficulty swallowing and slurred speech while maintaining an alert mental state (Table 4). There may not always be gastrointestinal symptoms. Symptoms progress via a characteristic descending bilateral flaccid paralysis to ventilatory paralysis and potentially, death (McLauchlin et al., 2006).

The incubation time of wound botulism is hard to determine as most patients will inject drugs multiple times per day, so it is difficult to know at what time exposure occurred. The neurological symptoms are indistinguishable from food-botulism, but there are no gastrointestinal symptoms. The abscess is often a minor lesion, resembling mild cellulitis (Sobel, 2005).

The initial symptoms of infant botulism are less apparent than food-borne botulism with the most common and earliest being constipation. There is usually a weak cry and general weakness, feeding difficulty, poor sucking, lethargy, lack of facial expression, irritability and progressive 'floppiness'. Ocular dysfunctions including ptosis and dilated and sluggish pupils usually become evident during the course of the disease. Respiratory arrests frequently occur but are rarely fatal due to access to infant intensive care medicine including artificial ventilation. Diagnosis is difficult due to lack of specific symptoms and variety of disease severity as well as the

rarity of disease. Identification of BoNT or *C. botulinum* in the stool is important for diagnosis (Dodds, 1992b; Lindstrom & Korkeala, 2006). Most infant botulism patients require hospitalisation and then return to full health, although there are a small number of extremes case where out patient care is sufficient or death occurs. Of the 2419 cases in the USA between 1976 and 2006, 9 patients (0.4%) did not require hospitalisation and 20 infections resulted in the death of the infant (0.8%) (Koepke et al., 2008). Mild cases are likely under-diagnosed.

Botulinum neurotoxin (BoNT) A is associated with more severe disease than BoNT/B or E (Woodruff et al., 1992). People intoxicated with BoNT/A consult clinicians earlier, are more likely to need artificial ventilation and require longer hospitalisation (Hughes et al, 1981). Recovery requires the growth of new neuromuscular connections and it is not uncommon for patients to require 2-8 weeks of ventilator support; whilst some patients may require several months of support before the return of muscular function (Shapiro et al., 1998). One factor behind the longer effects of BoNT/A compared with BoNT/E is that BoNT/E is ubiquitylated and rapidly degraded in cells while BoNT/A appears much more stable (Tsai et al., 2010).

Table 4: Symptoms of foodborne botulism caused by toxin types A and B*

Symptoms	Cases, %	Signs	Cases, %
Fatigue	77	Alert mental status	90
Dizziness	51		
		Ptosis	73
Double vision	91	Gaze paralysis	65
Blurred vision	65	Pupil dilation	44
		Nystagmus	22
Dysphagia	96		
Dry mouth	93	Facial Palsy	63
Dysarthria	84		
Sore throat	54	Diminished gag reflex	65
		Tongue weakness	58
Dyspnea	60		
		Arm weaknesss	75
Constipation	73	Leg weakness	69
Nausea	64	Hyporeflexia	40
Vomiting	59	Ataxia	17.
Abdominal cramps	42		
Diarrhea	19		
Arm weakness	73		
Leg weakness	69		
Paresthesia	14		

***Data is from the 1981 Hughes et al study on outbreaks of botulism reported in the United States in 1973-74, number of patients varied from 35 to 55.**

Botulism is a clinical diagnosis with laboratory confirmation by detection of BoNT and isolation of *C. botulinum*. Diagnosis of botulism is based on a combination of the presence of compatible clinical symptoms and laboratory confirmation through detection of BoNT in serum, faeces or suspect foods. Diagnosis and subsequent treatment should not depend on the results of laboratory testing due to the life-threatening nature of the disease and need for urgent treatment (Lindstrom and Korkeala, 2006).

The gold standard diagnostic test for botulism is the detection of BoNT by mouse bioassay (MBA) (Lindstrom and Korkeala, 2006). Faecal material or a suspected source of intoxication is homogenised, filtered (to prevent infection) and injected intra-peritoneally into mice. If the sample is positive for BoNT then the animals will show signs of intoxication within 1-4 days. If infection by a non-proteolytic strain is suspected the homogenised sample is treated with trypsin to activate the toxin. This is done as standard as it is not usually known whether an implicated organism is proteolytic or not. Neutralisation of classic symptoms in the MBA using specific anti-sera confirms the presence of BoNT and also determines toxin type. There are numerous issues with the MBA including the fact that it takes 4 days to return a negative result and the cost and ethically sensitive nature of animal testing. Also, the sensitivity of the MBA at identifying wound botulism infections is not ideal for all scenarios; one study found that only 50/73 (68%) wound botulism tissue samples were positive by MBA (Wheeler et al., 2009). Therefore molecular techniques have been widely adopted as a rapid, sensitive and specific test to supplement the MBA. Multiplex conventional polymerase chain reaction (Lindstrom et al., 2001) and real time polymerase chain reaction (Akbulut et al., 2004; Grant et al., 2009) assays have been developed for the detection of the

different subtypes of BoNT encoding gene. These assays can be used to screen suspect material/patient samples to confirm or reject a clinical diagnosis (K. Grant, personal communication).

In addition to direct toxin or molecular testing, the isolation and identification of the causative organism is important for reference microbiology and outbreak investigation. *C. botulinum* isolation medium uses a base of egg yolk agar supplemented with cycloserine, sulfamethoxazole and trimethorpim to select group I *C. botulinum*. The presence of the egg yolk agar enables the lipase reaction typical of *C. botulinum* (Lindstrom & Korkeala, 2006).

Symptomatic treatment, in particular the use of artificial respiration to alleviate respiratory failure, has been vital in reducing botulism mortality (Sobel, 2005). In addition to intensive care treatment, BoNT antitoxin is a widely used therapeutic. It binds and inactivates BoNT, thus preventing the toxin binding at the neuromuscular junction, to be effective antitoxin needs to be administered early on in the onset of disease, preferably < 24 h after onset of symptoms so it can deactivate circulating toxin before the BoNT enters the neuron (Sobel, 2005). Antibiotic treatment is not advised in treating food or infant botulism as lysis of any *C. botulinum* may lead to an increase in the amount of circulating toxin. However, metronidazole is often used to treat the wound botulism, after antitoxin treatment.

Until recently the only treatment for infant botulism was supportive therapy such as artificial ventilation. Equine antitoxin used for treatment in adult botulism is not advised due to the high risk of infants suffering a serious adverse reaction.

However, in 2003 an immune globulin treatment was licensed by the Food and Drug Administration in the USA and is now available from the Infant Botulism Treatment and Prevention Program at cost (£64000). The treatment is known as Botulism Immune Globulin Intravenous (BIG-IV or BabyBIG) and is produced from high-titre immune plasma donated by volunteers who have been immunized with pentavalent (A-E) botulinum toxoid. Treatment with BIG-IV has resulted in a total of 30 years of avoided hospital stay and saved more than $50 million in hospital costs. It cost approximately $10 million and took nearly 15 years to develop, license and produce BIG-IV (Arnon, 2007). Even after antitoxin administration, it is still vital that supportive therapy is applied to ensure a full recovery.

1.5. Botulinum Toxin

The botulinum neurotoxin (BoNT) is a bacterial exotoxin produced by *C. botulinum* and, less frequently, by other *Clostridium* species. It is widely regarded as the most toxic protein (Lamanna, 1959; Simpson, 2004). The potency of BoNT is such that no one has yet quantified the minimum concentration, or minimum number of molecules required to disrupt a vulnerable cell (Simpson, 2004) although it has been estimated that 1 gram of crystalline toxin would be sufficient to kill more than 1 million people (Arnon, 2001). This potency is all the more remarkable considering the long and complex route between the toxin point of origin (i.e. the gut in food-borne and infant botulism) and the final site of toxin activity (i.e. the pre-synaptic neuron cytosol) (Simpson, 2004).

A neurotoxin is a toxin that acts specifically on nerve cells, neurotoxins are among the most potent of natural and man-made toxins (Simpson, 2004; Munro et al, 1994). Botulinum and tetanus neurotoxins are orders of magnitude more toxic than any other naturally produced neurotoxins (Table 5).

Seven botulinum neurotoxin serotypes (types A-G) have been described with six of seven serotypes having subtypes (Montal, 2010). The different BoNT serotypes vary by 35-70% of their amino acid sequence while BoNT subtypes vary by between 2-32% of their amino acid residues (Peck et al, 2011).

Table 5: Potency of various natural neurotoxins (adapted from Gil, 1982).

Source	Test	LD_{50} µg/kg
Bungarus multicinctus, Taiwanese banded krait	Mouse, intraperitoneal LD 50	14
Oxyuranus microlepidotus, Australian taipan	Mouse, intravenous LD 50	2
Clostridium botulinum	Mouse, intraperitoneal LD 50	0.0012
Clostridium tetani	Mouse, unknown LD 50	0.001
Crotalus durissus cascavella, Brazillian rattlesnake	Mouse, intravenous LD 50	82
Naja haje, Egyptian cobra	Mouse, sub-cutaneous LD 50	50

1.5.1. Bacterial toxin complexes

Botulinum toxin is produced as part of a non-covalently associated, 300-900 kDa protein complex together with associated non-toxic proteins (ANTPs) (Sugii & Sakaguchi, 1975). The full contribution of the ANTPs to BoNT toxicity remains controversial (Simpson, 2004; Fujinaga, 2010).

Botulinum neurotoxin is not the only bacterial protein toxin produced as part of a hetero-multimeric toxin complex; the insect pathogens *Xenorhabdus nematophilus* and *Photorhabdus luminscens* also produce such complexes. These pathogens reside in the gut of entomophagous nematodes and are released into the insect's hemocoel (circulatory system) by the nematode where they grow, produce toxin and typically kill the insect with 24-48 h (Bowen et al., 1998). The toxins produced by these two species are heteromultimeric toxin complexes, known collectively as Tc family proteins (Waterfield et al., 2001). The toxicity and breadth of insect host susceptibility made the *P. luminescens* toxin complex a potential candidate for use in transgenic plants as a form of insect control (Bowen et al, 1998). The three part toxin complex is encoded by co-located genes (Waterfield et al., 2001) and was first identified in *P. luminescens* but has been more thoroughly characterised in *X. nematophilus*. In this species the three parts of the toxin complex are known as as XptA2 (280 kDa); XptB1 (170 kDa); and XptC1 (77 kDa) and bind together with a stoichiometry of 4:1:1 resulting in a toxin complex with an overall molecular weight of around 1 megadalton. The XptA2 tetramer binds to insect gut membranes and forms a pore in a model lipid bilayer system clearly implying a role in transport across the gut epithelium. XptB1 and C1 form a strongly associated complex which then binds to the tetrameric XptA2 forming the complete and fully active

toxin complex (Sheets et al, 2011). The class C protein component *is* an ADP-ribosyltransferase which modifies actin causing it to cluster resulting in disruption to the insect gut epithelium. The role of class B proteins is thought to be a chaperone or link between the A and C components (Lang et al, 2010). Orthologues of the toxin complex genes have been found in a range of bacterial pathogens associated with insects (e.g. *Yersiniae*). The *P. lumienscens* toxin complex shows that a large toxin complex can be essential in toxin activity (Sergeant et al., 2003).

Another bacterial toxin produced as part of a large protein complex is the pertussis toxin, produced by *Bordetella pertussis*, causative agent of whooping cough. Pertussis toxin (PTX) is one of the most complex soluble bacterial proteins and is composed of 5 subunits arranged in an AB structure (Tamura et al., 1982). The activity subunit consists of a single protein, an enzyme with an ADP-ribosylating activity. The binding subunit consists of the other 4 proteins that make up PTX, in a 1:1:2:1 stoichiometry (Stein et al., 1994). The heterologous binding subunit allows the PTX to bind to a variety of target cells, including almost all mammalian cells, with binding mediated by a variety of glycoproteins, glycolipids and receptor proteins. This promiscuity of binding may account for the wide spectrum of biological activities of the PTX (Locht et al., 2011).

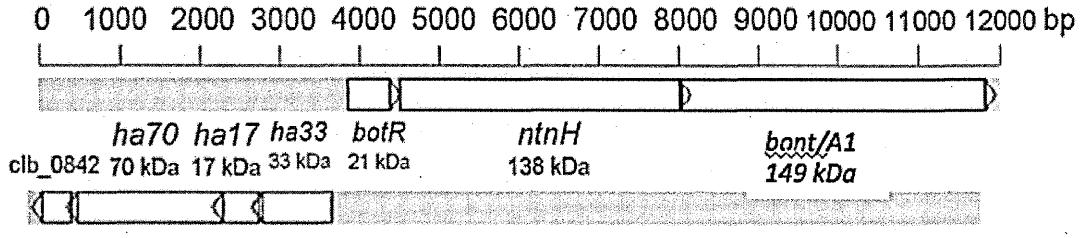
1.5.2. Botulinum toxin and the associated non-toxin proteins

Botulinum neurotoxin is encoded alongside and co-expressed with an associated non-toxic protein encoding gene cluster that is 12000 base pairs (bp) to 14000 bp in length, depending on toxin gene cluster subtype. There are two known subtypes

of toxin gene cluster – the haemagglutinin (HA) type (Henderson et al., 1996) and the OrfX type (Dineen et al., 2004). The HA type gene cluster consists of *bont, ntnH* and *botR* which are encoded on the forward strand and *ha33, ha17* and *ha70* which are encoded on the reverse strand (Figure 3A). The naming of these genes varies in the literature, this is the nomenclature that will be used here. In *C. botulinum* A1 ATCC 19397 *bont/A* is 3886 bp long encoding a 149.4 kDa protein, *ntnH* is 3581 bp long encoding a 138.2 kDa protein and *botR* is 536 bp long encoding a 21.7 kDa protein. In ATCC 19397 *ha33* is 881 bp long encoding a 33.8 kDa protein, *ha17* is 440 bp long encoding a 17 kDa protein and *ha70* is 1881 bp long encoding a 71.1 kDa protein.

The genomic arrangement of the OrfX type cluster (Figure 3B) differs from that of the HA cluster. The OrfX cluster consists, with some variants, of *bont, ntnH* and *p47* encoded on the forward strand and *botR, orfX/1, orfX/2* and *orfX/3* encoded on the reverse strand. In *C. botulinum* A3 Loch Maree *bont/A* is a 3879 bp gene encoding a 148.6 kDa protein, *ntnH* is 3480 bp long encoding a 134.6 kDa protein, *p47* is 1248 bp long and encodes a 47.4 kDa protein. On the reverse strand *botR* is 552 bp long and encodes a 22.2 kDa protein, *orfX/1* is 429 bp long and encodes a 16.5 kDa protein, *orfX/2* is 2253 bp long and encodes a 84.3 kDa protein and *orfX/3* is 1473 bp long and encodes a 55.2 kDa protein. BoNT/A1, A5, B, C, D and G have been identified encoded in HA gene clusters while BoNT/A1 (rarely), A2, A3, A4, E and F have been identified encoded in OrfX gene clusters. A schematic of the various arrangements of the HA and OrfX encoding gene clusters can be seen in Figure 4. BoNT/H, which has only recently been identified, in a bivalent Bh strain, is encoded as part of an OrfX cluster.

(A)



(B)

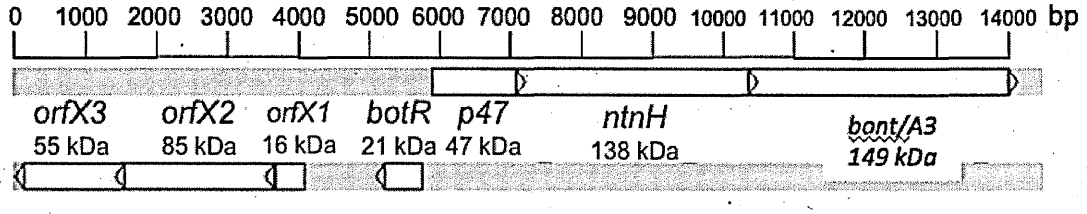


Figure 3: Genomic arrangement of (A) *bont/A1* in a HA gene cluster from *C. botulinum* A1 ATCC 19397 and (B) *bont/A3* from *C. botulinum* A3 NCTC 2012 in an OrfX gene cluster.

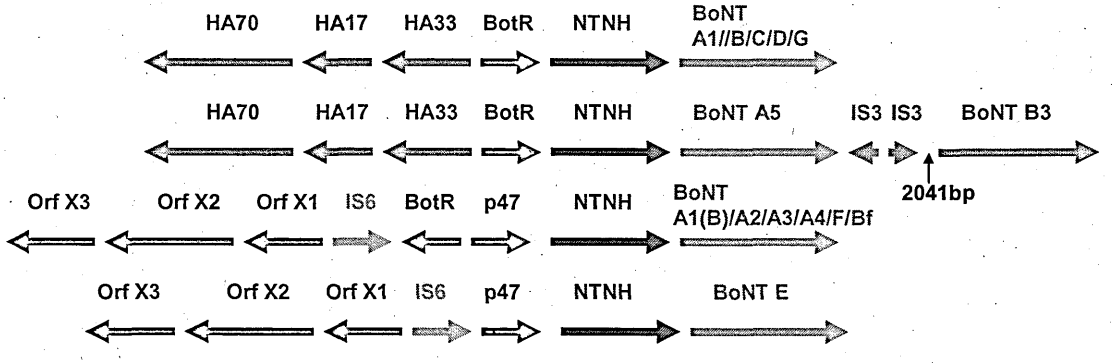


Figure 4: The different neurotoxin cluster arrangement in HA+ve/OrfX-ve strains, HA-ve/OrfX+ve strains and the bivalent A5b strain.

The BoNT-HA complex ranges from 300-900 kDa depending on which ANTPs form the toxin complex. These different complexes are known as 12S, 16S and 19S and the ANTPs present in each one can be seen in Figure 5. Multiple different complex sizes can be identified within the same culture and their relative abundance depends on strain, serotype and growth conditions (Sugii & Suguyama, 1975).

The toxin complexes of strains encoding BoNT/A2-OrfX have been examined (Lin et al., 2010). They only purified one 300 kDa BoNT complex, composed of BoNT/A2 and NTNH, with no OrfX cluster proteins. Previous work by the same group had found that *bont, ntnH* and *p47* were transcribed on the same tri-cistronic mRNA (Dineen et al., 2004). Other workers examining the complexes formed by a BoNT/E producing strain found a spontaneous association between BoNT/E and the OrfX proteins (Kukreja & Singh, 2007).

A recent paper has shed more light on the role of the HA toxin complex in the toxicity via the oral route of the botulinum toxin (Lee et al., 2013). They reported that the BoNT 16S protein complex consisted of 14-subunits with a combined molecular weight of ~760 kDa (**Figure 6**). They proposed a model whereby the 16S protein complex is divided into two structurally and functionally independent sub-complexes. One that consists of BoNT-NTNH while the other consists of HA70:HA17:HA33 in a 3:3:6 stoichiometry (Lee et al., 2013).

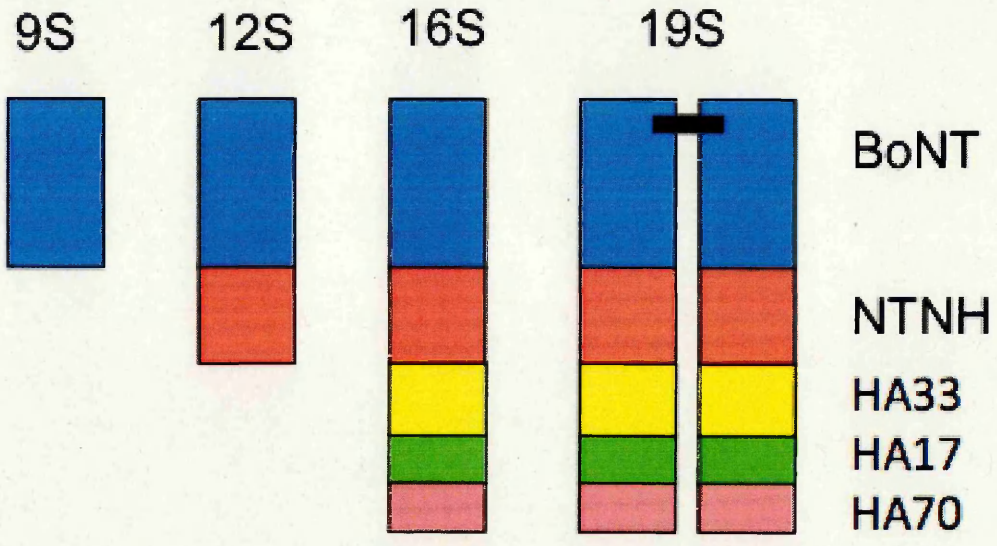


Figure 5: The different compositions of the complexes known as 9S, 12S, 16S and 19S. BoNT is Botulinum Neurotoxin, NTNH is Non-Toxic Non-Haemagglutinin, HA33, 17 and 70 have haemagglutinin properties. NTNH and HA33, 17 and 70 are Associated Non-Toxic Proteins (ANTPs).

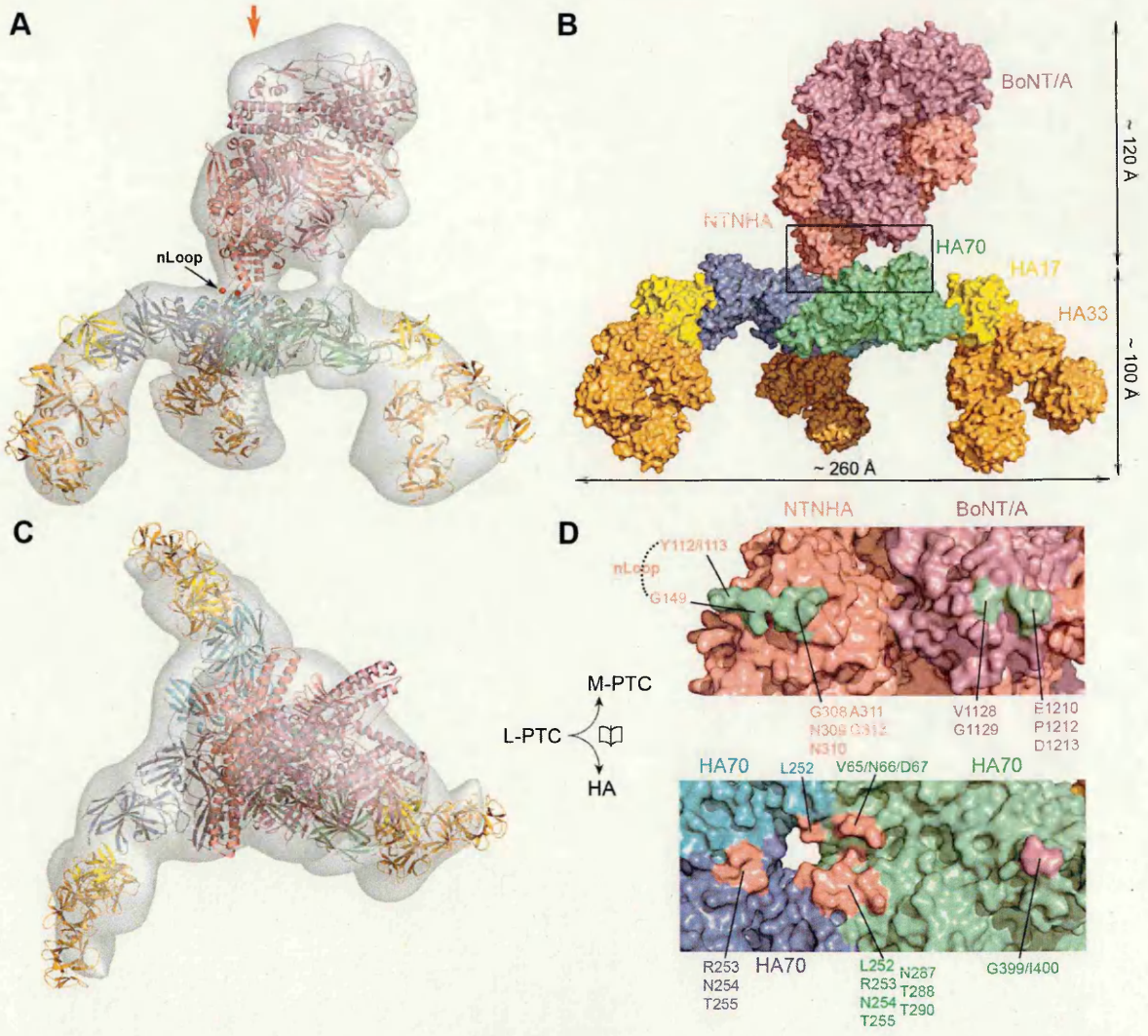


Figure 6: The molecular architecture of the large BoNT/A toxin complex (A) 3D-EM reconstruction of large BoNT/A toxin complex. The red arrow indicates the viewing direction of (C). (B) Surface representation large BoNT/A toxin complex in the same orientation as (A). (C) A different view of large BoNT/A toxin complex. (D) An open-book view of the interface that is highlighted in the box in (B)

1.5.3. The role of the ANTPs in BoNT intoxication

The toxicity of BoNT via the oral route is increased 100-1000 fold by the presence of associated non-toxic proteins (ANTPs) but the full role of the ANTPs in the intoxication of a botulinum poisoning victim has not been elucidated (Ohishi et al., 1977; Sakaguchi, 1982; Kukreja & Singh, 2007). The ANTPs bind non-covalently to the toxin forming a heterogenous complex of 290-900 kDa. The oral toxicity of BoNT increases with incremental association with ANTPs. Type B toxin shows this effect most dramatically with 12S toxin 20 times more toxic than BoNT alone and 16S 1000 times more toxic than 12S (Sakaguchi, 1982). This increase is, at least, partially explained by the close interlocking interaction of BoNT and NTNH resulting in protection against acids and proteases in the gut (Gu et al., 2012).

Botulinum toxin proteins are too large to be efficiently taken up in the gut up by paracellular diffusion. Toxin uptake through the gut epithelium occurs by a process of receptor binding to epithelial cells with the upper small intestine being the most important site of toxin uptake in both infant and food botulism (Fujinaga, 2010). At present there are two schools of thought as to how the toxin crosses the gut epithelium. The main difference between them is whether the moiety responsible for toxin binding to the gut epithelium is on the toxin or toxin complex. Simpson and colleagues, have argued that the binding capability is mediated primarily by the toxin itself. The counterview, held by Oguma and associates, is that the binding domain is present on the toxin complex.

In vitro experiments have shown that in the absence of ANTPs BoNT/A and BoNT/B bind to polarized human intestinal epithelial cells and undergo transcytosis from the apical to basolateral side (Maksymowych & Simpson, 1998). The binding site on the toxin is proposed to be on the carboxy-terminal portion of the heavy chain, which is also the location of the neuronal cell-binding site (Maksymowych & Simpson, 2004). Another parallel with binding to neuronal cells is that gangliosides, including GD_{1b} and GT_{1b}, and the SV2 protein have been implicated in the transcytosis of BoNT/A through polarised human epithelial cell lines (Caco-2 or T84 cells) (Couesnon et al, 2009). Once bound, BoNT/A is transported through polarized human intestinal epithelial cells in 30-60 minutes. No significant difference in transcytosis was found between BoNT/A in pure and complexed states after 120 minutes when assayed by mouse bioassay. BoNT/A passage rate was 10-fold more efficient through an intestinal crypt cell line (m-IC_{cl2}) which has a higher level of SV2 proteins than Caco-2 cells (Couesnon et al, 2008). These studies found no significant increase in toxin transcytosis if the ANTPs were present.

A different mechanism was proposed following experiments using ligated intestinal loops of guinea pigs and type C toxins which showed that only when toxin was complexed with the NTNH and HA proteins did it bind the microvilli of the upper small intestine via sialic acid residues in cell surface glycoconjugates (Fujinaga et al, 1997). Type C toxin in complex with NTNH and HA proteins was shown to bind and enter the human intestinal epithelial cell line (HT-29) via cell surface sialic acid containing glycoproteins. This binding activity was much less significant in the absence of the HA proteins (Nishikawa et al, 2004) leading to speculation that the HA proteins may play a pivotal role in the absorption of BoNT from the lumen of

the gut. Studies on type D toxin also implicated the HA proteins in having a key role in transcytosis of this serotype (Niwa et al, 2007).

These sets of experiments propose directly opposite ideas. However, most of the work was done using in vitro cell lines that are not always physiologically relevant. In terms of the initial binding site of the BoNT complex to the gut epithelium neither BoNT nor the ANTPs can be ruled out and the true situation is likely to be a combination of both. There may also be subtype specific modes of toxin binding to the gut epithelium, as there are for binding of the toxin to neuronal cells.

Evidence has recently emerged which suggests that the role of the ANTPs may be involved in more than just binding; HA from BoNT/B has been found to disrupt the epithelial barrier by opening intercellular tight junctions (Matsumara et al, 2008). When 16S toxin complex was added to the apical side of Caco-2 cells there was a time-dependent decrease in transepithelial resistance (TER, a measurement of the 'intactness' of a membrane) with almost complete abolishment by 48 h (Matsumara et al, 2008). This epithelial disruption resulted in 10-fold more BoNT/B crossing Caco-2 monolayers after 24 h when toxin complex was used rather than just BoNT or BoNT/NTNH. The epthelial disruption may have been dismissed as non-specific deterioration of the in vitro membrane in previous experiments. The TER disruption was not observed when buffer, pure toxin or toxin complexed with NTNH was applied. TER was maintained if the toxin complex was treated with anti-HA antibodies before application to cells while disruption still occurred if the toxin complex was treated with anti-BoNT antibodies prior to application (Matsumara et al., 2008).

The disruption of the epithelium occurs from the basolateral side of the gut epithelium. When the toxin was added basolaterally both the 16S toxin complex and the HA proteins alone caused abolishment of TER within a few hours rather than 48 h when applied apically. No cell death was detected indicating a specific mode of action in the abolishment of TER (Matsumara et al., 2008). The in vivo effect was measured by assaying the intestinal absorption of an inert fluorescein dextran for which no active take up mechanism exists in the gut (Matsumara et al., 2008). When mouse intestinal loops were treated with ANTPs uptake of fluorescein dextrans of 4 kDa and 10 kDa were strongly enhanced, uptake of a 150 kDa fluorescein dextran was weakly enhanced. The ANTPs were intraintestinally co-injected with BoNT/NTNH into mice reducing the time until death two-fold. As the ANTPs were not complexed with the BoNT/NTNH this effect was not due to protection from digestion. Thus the ANTPs seem to compromise the intestinal barrier function facilitating the absorption of BoNT and other proteins by passive diffusion (Matsumara et al, 2008).

Further work by the same group revealed that BoNT/B HA disrupts epithelial cadherin (E-cadherin), a calcium dependent cell adhesion protein involved in tight end junction binding (Suguwara et al., 2010) (Figure 7). There is variation between animal species in their susceptibility to different BoNT types. This variation could be due to the specificty of interaction between the products of HA genes encoded alongside those toxin types and the epithelium protein E-cadherin (Suguwara et al., 2010). BoNT/B HA interacts with human, mouse and bovine E-cadherin but not rat or chicken. Avian botulism caused by type B strains is rare and birds have been shown to be resistant to BoNT/B, especially when administered orally (Gross &

Smith, 1971). HA from BoNT/C encoding strains, which don't cause human botulism, is unable to interact with human E-cadherin (Suguwara et al., 2010).

Several pathogens (*Bacteroides fragilis*, *Porphyromonas gingivalis* and *Candida albicans*) produce proteases that cleave E-cadherin, leading to epithelial barrier disruption and tissue invasion. BoNT/B HA also disrupts the epithelial barrier in an E-cadherin mediated way but without proteolytic cleavage – the precise mechanism remains unknown (Sugawara et al, 2010).

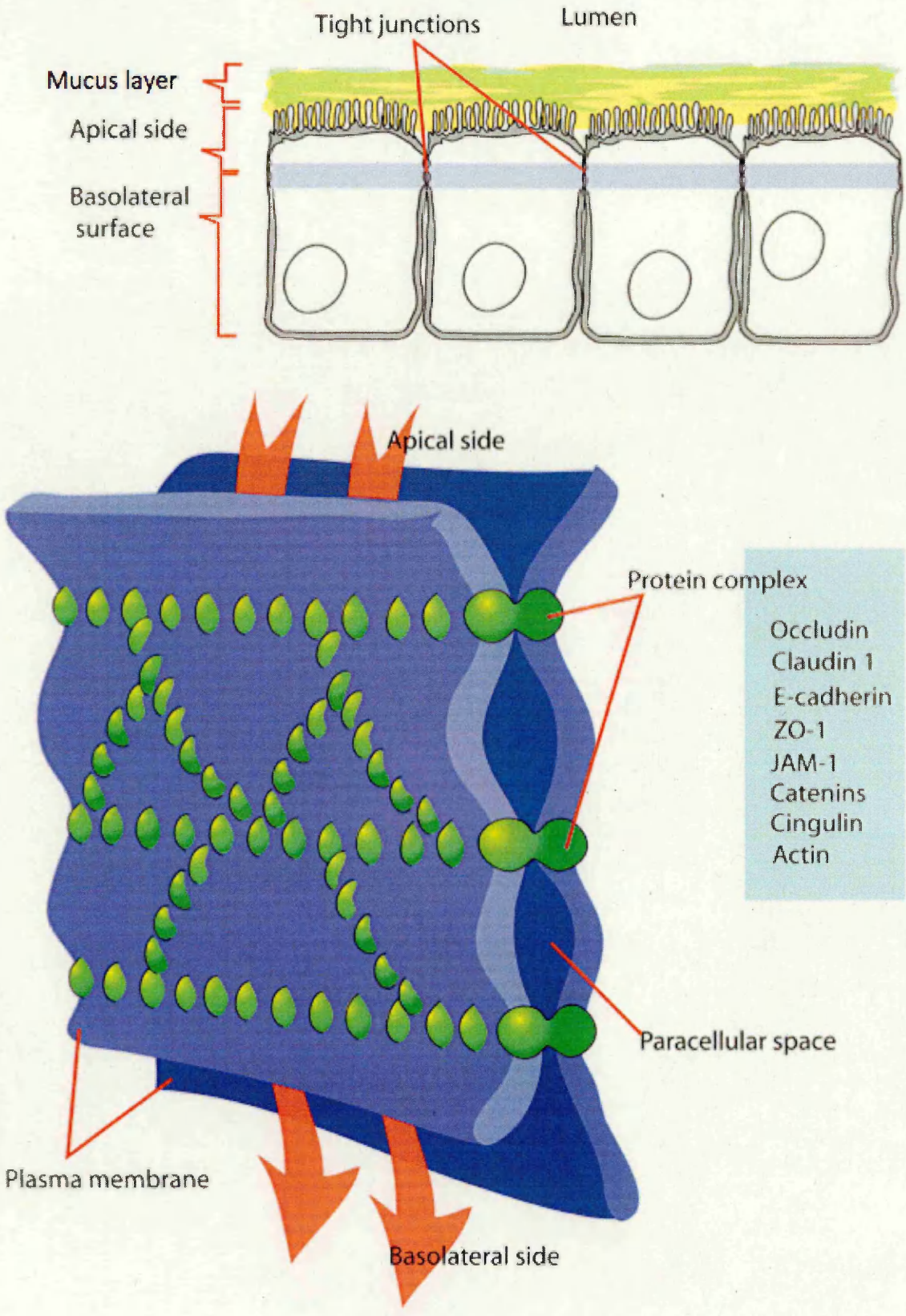


Figure 7: BoNT/B disrupts epithelial tight junctions via an E-cadherin dependent mechanism. A diagram of the location of tight junctions within the epithelial barriers (top) and a schematic of the tight junction itself (bottom). The protein complexes responsible for tight junction integrity are shown in green and include E-cadherin which is key to the integrity of the tight junction (adapted from Gruenheid et al, 2003).

These discoveries led to a 3-step model for BoNT crossing the gut epithelium being proposed. Step 1 is transcytosis mediated by either HA or BoNT heavy chain, step 2 involves basolateral HA disrupting E-cadherin, damaging the gut epithelium and facilitating step 3, the non-specific paracellular movement of macromolecules, including BoNT to the basolateral side of the gut epithelium (Fujinaga, 2010). This model portrays the BoNT complex as a multifunctional protein assembly equipped with the machinery to efficiently breach the intestinal barrier enabling the toxin to act on peripheral Cholinergic cells.

As stated previously the toxin can be produced as part of two different toxin complexes, the HA type and the OrfX type. The previous paragraphs describe the HA complex as most research has been done on this type due to the fact that most isolates from clinical cases, particularly of food botulism, have been found to be caused by HA encoding strains. The toxin complex type is rarely determined for clinical cases of botulism, as it is not deemed clinically relevant. However, investigating the distribution of HA and OrfX toxin complex types among clinical strains from different forms of botulism may provide insight into the differences between the two toxin complex types and any associations with food, wound or infant botulism.

Insight into the possible role of the ANTPs can be gained by comparing the situation with the botulinum neurotoxin with that of the related tetanus neurotoxin (TeNT). Two striking dissimilarities between BoNT and TeNT are that TeNT lacks a toxin complex and is incapable of poisoning via oral ingestion. The fact that TeNT causes intoxication through infected wounds indicates that ANTPs are not

necessary for this to occur (Singh et al, 1995). BoNT and tetanus neurotoxin are closely related proteins with approximately 30% amino acid identity and similar modes of action (Montecucco & Schiavo, 1994).

Recent work has been done on the structure and function of the BoNT complex (Lee et al., 2013). It was found that the hetero-dodecameric, 470 kDa complex (HA70:HA17:HA33 in a 3:3:6 stoichiometry) facilitates the absorption through the gut epithelium. The absorption is mediated by nine glycan-binding sites on the HA sub-complex that form multivalent interactions with carbohydrate receptors on intestinal epithelial cells. Competitive antagonists of these carbohydrate receptors blocked intoxication via the oral route in mice (Lee et al., 2013). It was also found that the disruption of the gut epithelium by the HA sub-complex was more effective when applied from the basolateral side than the apical side, mirroring the findings of Sugawara et al. (2010).

1.5.4. Major steps in botulinum neurotoxin action

Botulinum neurotoxin is synthesised as a single-chain, 150 kDa polypeptide. The precise mechanism by which the toxin is released from the bacterial cell remains unelucidated. Botulinum neurotoxin does not have a signal peptide indicating that it is secreted by a non-signal peptide secretion system such as the flagella export apparatus (Rao et al, 2007).

In order for BoNT to inhibit cholingeric transmission it must first be activated. This activation takes the form of host or endogenous proteases nicking the single chain polypeptide produced by the bacteria to create a dichain consisting of a heavy

chain (around 100 kDa) and a light chain (around 50 kDa) which are linked by a disulphide bond (Simpson, 2004; Barth et al., 2004).

The toxin complex spontaneously disassociates at the physiological pH of 7.4 and ionic strength it is exposed to after it has crossed the gut epithelium (Simpson, 2004). Once BoNT has crossed the gut epithelium it enters the circulation and then has to exit the vasculature and enter the extracellular space in the vicinity of the peripheral cholinergic nerve ending. However, the mechanism for this egress is unknown with no studies performed on this aspect of BoNT action (Simpson, 2004). It is known that large molecules can diffuse between vascular endothelial cells which raises the possibility that some of the toxin egress is due to passive diffusion. The fact that BoNT cannot pass across the much less permeable blood brain barrier lends a certain amount of strength to this hypothesis. However, much work needs to be done before any certainty can be accorded to this idea and such a non-specific mechanism seems unlikely for account for the transfer a potent toxin (Simpson, 2004).

Once in the vicinity of the pre-synaptic cholinergic nerve cell, BoNT is taken up by a four-step process before exerting its effect, namely:

1. Binding to the plasma membrane
2. Receptor mediated endocytosis
3. pH-induced translocation across the endosome membrane
4. Intracellular cleavage of target protein

The first step, the binding of BoNT to nerve cells, occurs according to a dual receptor model involving two types of receptor; membrane gangliosides and synaptic vesicle proteins.

BoNT first binds to gangliosides that are enriched in neuronal plasma membranes, specifically disialo- and trisialo- gangliosides (GD1b, GT1bs), with affinities in the nM range (Montal, 2010). BoNT/D is the only sub-type that hasn't been shown to bind gangliosides but it binds phospholipid phosphatidylethanolamine instead (Montal, 2010). One proposed model involves a two step process; the initial binding of the toxin is mediated by gangliosides that increases the toxin concentration in the vicinity of the plasma membrane, the toxin can then diffuse in the plane of the membrane, bind it's protein receptor and be endocytosed (Stenmark et al, 2008). BoNT crosses the neuronal cell membrane by essentially, hitchhiking with a vesicle that is being recycled (Figure 8). The neuronal membrane protein receptor for BoNT/A, E and F is the human synaptic vesicle protein SV2 (Synaptic Vesicle protein 2) (Dong et al., 2006; Montal, 2010). BoNT binds to the SV2 on the luminal surface of the synaptic vesicle and is then endocytosed when the vesicle is recycled. This relationship between synaptic vesicle release and BoNT/A entry also means that the most active nerve cells are intoxicated first (Dong et al, 2006). BoNT/B and G employ a similar mechanism but binding to different luminal secretory vesicle proteins, synaptotagmin I and II (Dong et al, 2003). It is thought that BoNT/C only binds to gangliosides and BoNT/D only binds to phosphatidylethanolamine and that neither have a protein receptor (Tsukamoto et al, 2005).

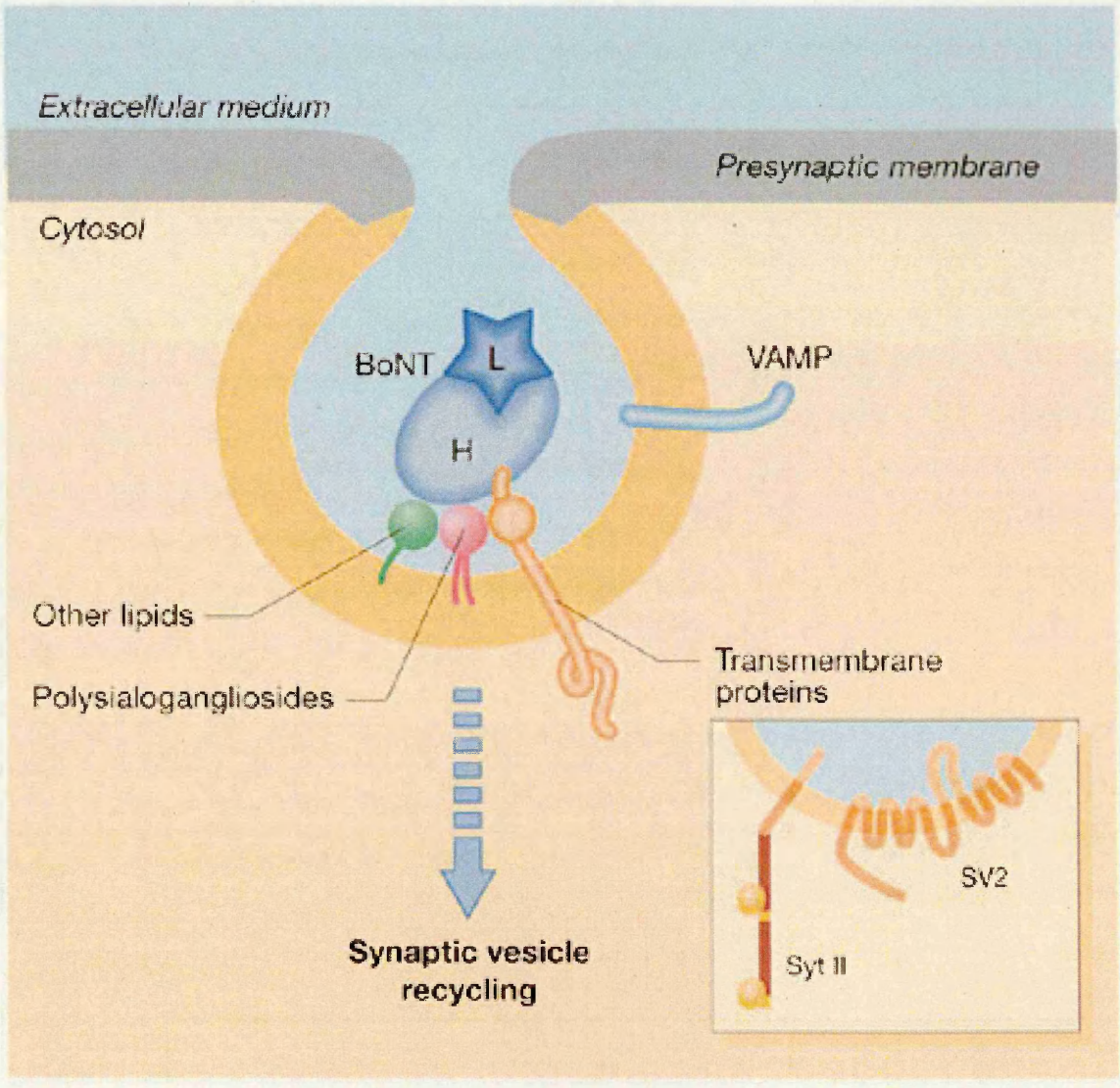


Figure 8: the binding of BoNT to the luminal surface of a recycling synaptic vesicle. The illustration depicts the combined role of polysialogangliosides (purple) and a synaptic vesicle (SV) protein (SV2 or synaptotagmin; orange) in mediating botulinum neurotoxin (BoNT)-neurospecific binding and entry into neurons after the retrieval of the vesicle. It suggests additional low-affinity interactions with other molecules (lipids and/or proteins) of the SV membrane (green; modified from Montecucco *et al.* 2004). Multiple interactions with molecules of the SV membrane would allow almost irreversible neuron binding, which would then become completely irreversible on vesicle fission from the presynaptic membrane. The yellow and grey areas denote the different compositions of the SV and presynaptic membranes, respectively. H, heavy chain of BoNT; L, light chain of BoNT; VAMP, vesicle-associated membrane protein. Adapted from Verderio et al., 2006.

In order to disrupt it's cytosolic target, botulinum neurotoxin must escape the endosome (i.e. the recycled vesicle). This occurs in a pH dependent manner similarly to diphtheria and other bacterial toxins (Simpson, 2004). The normal acidification of the endosome has two noteable effects on BoNT, the first of which is that it causes a major conformational change in the heavy chain. This conformational change results in the heavy chain being inserted into the endosomal membrane where it acts as a transmembrane channel/chaperone, a dynamic structural device that achieves translocation of the light chain through the endosomal membrane (Koriazova and Montal, 2003). The heavy chain also prevents aggregation of the light chain that would otherwise occur under acidic conditions. The second effect is a partial unfolding of the light chain allowing it to pass through the channel formed by the heavy chain (Fischer and Montal, 2007). If the toxin is not nicked (i.e. is a single polypeptide rather than two polypeptides linked by a disulphide bond) then the channel can form but is occluded by the light chain that remains in the vesicle. The light chain and membrane inserted heavy chain dissociate when the disulphide bond linking the two is reduced by the naturally reducing conditions in the cytosol. Once in the neutral pH of the cytosol the light chain refolds, regaining its active state, allowing the light chain to cleave its target substrate. In summary BoNT egress from the endosome requires acidic pH, non-reducing conditions inside the endosome and a neutral pH, reducing conditions in the cytosol (see Figure 9) (Koriazova and Montal, 2003; Fischer and Montal, 2007).

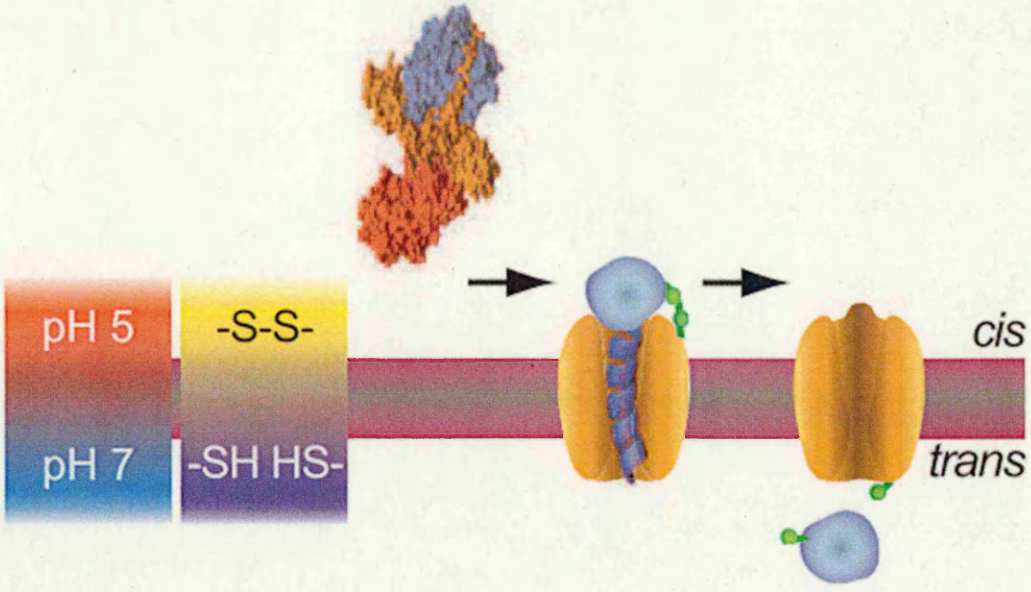


Figure 9: BoNT translocation through the endosomal membrane. Structure of BoNT/A (light chain = purple; translocation domain = orange; membrane binding domain = red) before insertion into the membrane (grey bar with magenta boundaries) and then a schematic of the membrane inserted BoNT/A during translocation of the light chain (purple) through the transmembrane domain of the heavy chain (orange) and the reduction of the disulphide bond (green). Acidic pH in the endosome and reducing conditions in the cytosol are shown on the left. Adapted from Fischer and Montal, 2007.

The light chain of botulinum neurotoxin is a zinc dependent endopeptidase that cleaves SNARE (Soluble NSF (N-ethylmaleimide sensitive fustion protein) Attachment Protein Receptor) proteins. The primary role of SNARE proteins is to mediate vesicle-plasma membrane fusion; in this case synaptic vesicles with the neuronal membrane (Schiavo et al, 1992; Blasi et al, 1993). Different subtypes of BoNT cleave different SNARE proteins. BoNT/A and E cleave SNAP-25, BoNT/C cleaves SNAP-25 and syntaxin while BoNT/B, D, F and G cleave vesicle associated membrane protein (VAMP), also known as synaptobrevin (Simpson, 2004). Furthermore each BoNT will only cleave a single peptide bond in it's substrate protein even if that peptide bond is repeated elsewhere in the substrate. It has been shown that the light chain requires quite large fragments of the target protein for efficient cleavage and single amino acid changes that are not in the immediate vicinity of the cleavage site can dramatically reduce proteolysis efficiency (Brunger and Rummel, 2009). This specificity is thought to be due to the presence of a common motif in the substrate SNARE proteins that acts as binding sites for the toxin (Rossetto et al, 1994). This common substrate motif present in all the BoNT target proteins explains the targeting of these varied proteins by similar enzyme. It also results in only one of the multiple identical peptide bonds being exposed to the toxin's active site, which explains why the SNARE protein is only cleaved at a single peptide bond (Simpson, 2004).

The duration of intoxication i.e. presence of clinical symptoms, of BoNT varies with serotype, BoNT/A has the most sustained toxicity which can last months while BoNT/E's toxicity is shorter, typically only several weeks (Elopra et al, 1998; Tsai et al., 2010). One explanation for this discrepancy is that different BoNT serotypes are degraded at different rates by ubiquitin mediated proteasomal degradation

within the neuron. BoNT/E light chain is preferentially targeted by the ubiquitinlyation machinery of the neuron when compared to BoNT/A light chain (Tsai et al, 2010). This leads to two interesting potential applications; the development of a recombinant BoNT for clinical use which is even more resistant to ubiquitination than BoNT/A and the development of a ubiquitin ligase targeted to BoNT/A light chain for use in cases of botulism (Tsai et al, 2010).

An informative perspective can be gained on the BoNT complex if it is compared with the tetanus neurotoxin (TeNT). TeNT has 30% amino acid identity with BoNT/A1, is a zinc dependent metallopeptidase which enters nerve cells and blocks neurotransmitter release by cleaving part of the cellular machinery involved in neuroexocytosis (Montecucco et al, 1994). There are differences in the precise mechanism of action at the nerve ending; TeNT undergoes retrograde axonal transport, being discharged into the synaptic cleft and then taken up by the post-synaptic cell where it cleaves VAMP (a SNARE protein, cleaved by B, D, F and G). It shares the three domain architecture of BoNT with the C-terminus of the heavy chain involved in neurone binding, N-terminus of the heavy chain involved in pH dependent release from the endosome and the light chain having enzymatic activity which results in the cleavage of VAMP.

Many bacterial toxins such as Shiga toxin, diphtheria toxin and anthrax toxin are AB toxin complexes. These consist of two proteins, the A protein which is responsible for the activity of the toxin and the B protein which is responsible for the binding and internalisation of the toxin to its specific target (Barth et al., 2004).

Botulinum neurotoxin is not thought of as an AB toxin as it is encoded by a single gene, however, it shares many characteristics with the AB toxins (see Figure 10).

Botulinum toxin, with it's light and heavy chains shares some characteristics with AB toxins. However, the presence of the ANTPs in addition to the AB characteristics of BoNT highlight the additional complexity of BoNT compared with AB toxin complexes. This could be due to the longer route of required for BoNT toxicity.

Although the *P. luminescens* toxin requires it's associated proteins for toxicity it is essentially a more complex variant of the binary AB toxin model. The XptA tetramer is the binding domain, XptC is the activity domain while XptB is likely a chaperone to facilitate interaction between the XptA and the XptC domains. The essential difference between the BoNT complex and any other toxin complex in the literature is that BoNT, in addition to having the A/B subunit architecture within the BoNT protein has large, highly conserved, associated non-toxic proteins the functional and epidemiological consequences of which have not yet been fully understood.

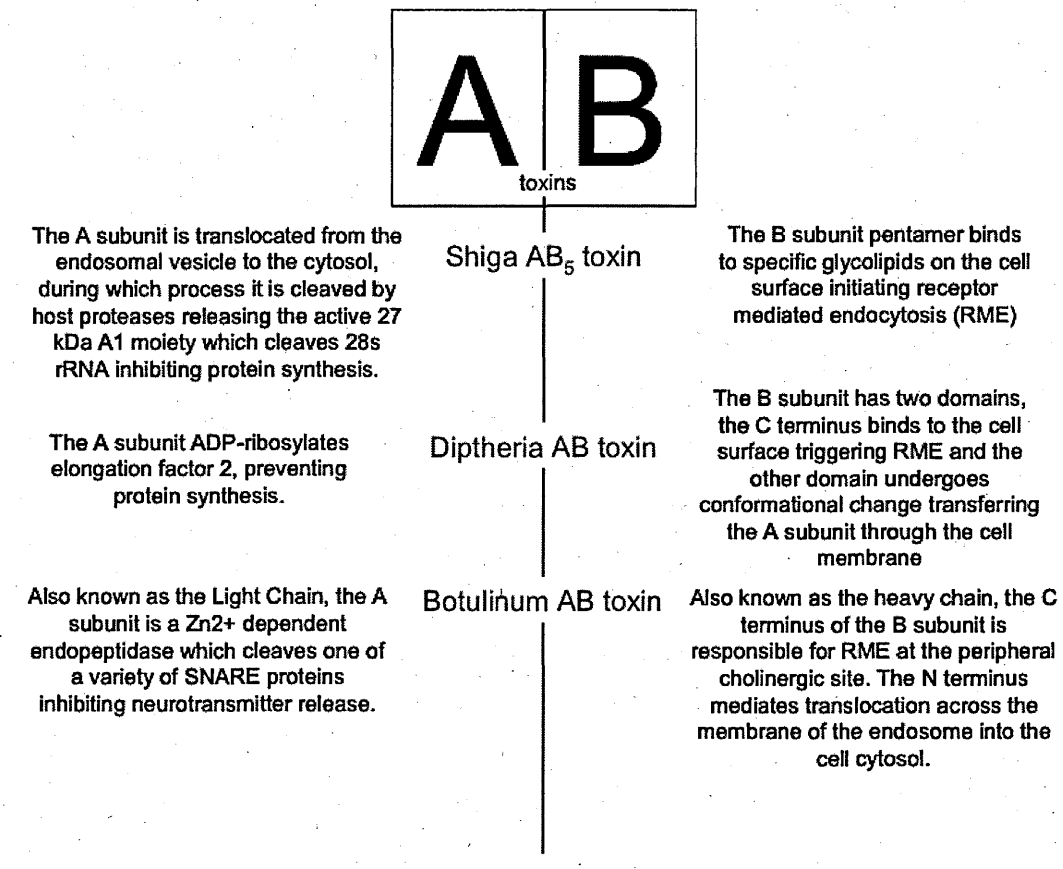


Figure 10: A comparison of Shiga and Diptheria AB toxins with botulinum toxin which is not classed as an AB toxin but which shares certain characteristics with AB toxins.

1.5.5. Evolution of BoNT and the associated non-toxic proteins

The presence of the botulinum toxin-encoding gene in distinct genetic backgrounds (i.e. group I-IV *C. botulinum*, *C. butyricum, C. beijerincki*) indicates that BoNT has likely undergone horizontal gene transfer multiple times in evolutionary history (Skarin & Segerman, 2011). The presence of the neurotoxin gene clusters on plasmids, bacteriophage and partial and incomplete insertion sequences flanking *bont* also lends weight to this hypothesis (Hill et al, 2009; Peck et al, 2011).

Clostridial neurotoxins (BoNT and TeNT) are only found in the genus *Clostridium* and possess a unique sequence and structural architecture compared to other protein families. Looking for distant homologues of BoNT in the *C. botulinum* genome found evidence that suggests that the neurotoxin evolved from an ancestral collagenase-like gene (Doxey et al, 2008). However, certain BoNT features such as autocatalysis, polyprotein architecture (meaning one gene encodes multiple proteins with separate functions), encoding on bacteriophage and similarities of the light chain to viral metallopeptidases indicate a possible viral evolutionary history (DasGupta, 2006).

It is likely that BoNT/A1 was originially associated with an OrfX type cluster while BoNT/B was associated with the HA type cluster. Then, a recombination event occurred approximately in the centre of the *ntnH* gene resulting in BoNT/A1 being combined with an HA type cluster. The resulting BoNT/A1-HA gene cluster contains an *ntnH* gene, the 3' end of which is similar to an OrfX *ntnH* and the 5' end of which is similar to an HA *ntnH* (Hill et al., 2009). This BoNT/A1 HA

encoding lineage is frequently associated with foodborne botulism (Peck et al, 2011).

1.5.6. Regulation of *bont* and the *Clostridium botulinum* transcriptome

One of the characteristic features of the pathogenic clostridia is their production of toxins. The structure, mode of action and encoding genes of the main clostridial toxins are well understood, however, the regulation of synthesis of these toxins has not been fully elucidated. The first clostridial toxin regulation system to be investigated and understood was that of the *C. perfringens* toxins which was found to be regulated by a two-component signal transduction system, VirR/VirS (Ba-Thein et al, 1996). VirS is a membrane-associated sensor that responds to an unknown signal by initiating a phosphorylation cascade resulting in phosphorylated VirR which then modulates transcription of its target genes. This modulation can be either by direct binding to promoters or by promoting transcription of a regulatory RNA molecule which in turn controls gene expression. In *C. perfringens*, the *VirR/VirS* system was found to be a global regulator responsible for the up-regulation and down-regulation of various genes including toxin genes, toxin gene regulators and amino acid synthesis genes (Banu et al, 2000). Homologues of VirR have been found in *C. difficile*, *C. botulinum* and *C. tetani*. Recent work has found that there are three two-component systems that, when knocked down, directly (i.e. with no effects on growth rate) induced a low level of BoNT/ANTP expression (Connan et al., 2012). This reduction in expression was found to be independent of *botR* and indicates that BoNT synthesis is under the control of a complex network of regulation. Interestingly, the two Two-Component Systems in

C. botulinum that show similarity to VirR were not directly involved in BoNT regulation (Connan et al., 2012).

Regulation of transcription of the gene cluster encoding BoNT and the ANTPs is thought to involve BotR (Raffestin et al., 2005), an alternative sigma factor that is a positive transcriptional regulator. The gene encoding BotR is in the same orientation as the *ntnH/bont* cluster in the HA encoding strains and in the same orientation as the *orfX* genes in the OrfX encoding strains. (Dupuy & Matamouros, 2006; Marvaud et al, 1998; Peck et al, 2011). No *botR* homologue has been identified in *C. botulinum* type E toxin gene clusters. Additionally, *C. botulinum* producing toxin type A5 does not have -35 and -10 binding sites upstream of the *botR* gene, this deletion has no noticeable adverse effect on BoNT production by this strain (Peck et al, 2011; Carter et al, 2010). This indicates that an alternative mechanism is involved in toxin gene regulation. There is a homologue of *botR* in *C. tetani* (*tetR*) that is thought to play a similar role, the signal which leads to activation of BotR/TetR is unknown. However, a homologue of the *Staphylococcus aureus* quorum sensing system (*agr*) has been found in all group I strains. When *agr* was knocked out, sporulation and toxin production were significantly reduced indicating that quorum sensing could be involved in toxin production (Cooksley et al, 2010). In another study, when these genes were knocked out there was no impact on toxin production leaving the role of quorum sensing in toxin production remaining to be elucidated (Connan et al., 2012).

Previous work on the *C. botulinum* transcriptome has focussed on the expression of the toxin gene. Work by Bradshaw et al in 2004 investigated the kinetics of

botulinum neurotoxin production in three strains of *C. botulinum* in two different media types from 0-96 hours. This interesting and thorough work used northern blotting to analyse mRNA levels and ELISA, Western blots and mouse bioassay to assay neurotoxin concentrations. In all three strains studied, mRNA transcripts for the toxin complex genes were initially detected in early log phase, reached peak levels in early stationary phase and rapidly decreased in mid to late stationary phase and during lysis. Toxin expression varied depending on strain and growth medium, highlighting the complexity of toxin regulation in *C. botulinum*. This difference between the time of peak toxin mRNA and peak toxin protein levels raises interesting questions about the dynamics of neurotoxin mRNA handling and toxin protein synthesis.

More recently, Artin et al. (2008) studied the effects of CO_2 on neurotoxin gene expression in non-proteolytic *C. botulinum* type E. they used reverse transcription quantitative PCR (RT-qPCR) and enzyme linked immunosorbent assays to quantify expression of the type E botulinum neurotoxin gene and formation of type E neurotoxin. Relative expression of the neurotoxin peaked in the transition between exponential phase and stationary phase then rapidly declined in stationary phase. The neurotoxin gene mRNA half-life was calculated to be approximately 9 minutes. Toxin protein formation occurred in late exponential and stationary phase. High CO_2 concentrations increased lag time and decreased the maximum growth rate. As a follow up to this work, Artin et al (2010) described the effects of carbon dioxide on the growth of proteolytic *C. botulinum* and the regulation of neurotoxin and the wider transcriptome. They used a combination of reverse transcription quantitative PCR (RT-qPCR) and a whole genome microarray. They found that CO_2 concentration (10%, 35% and 70%) had no

significant effect on gene expression or neurotoxin formation. There was also no significant effect on the growth curve. At all CO_2 concentrations the relative expression of neurotoxin cluster genes peaked in the transition between exponential and stationary phases with some evidence of a second rise in expression in late stationary phase. At 10% CO_2 their gene expression microarray found that neurotoxin gene expression was higher than the neurotoxin regulator (*botR*) as well as toxin complex components *ntnH, ha33* and *ha70* throughout the whole growth curve. They did however confirm previous findings that expression of neurotoxin cluster genes is growth phase dependent. They also showed that the neurotoxin cluster genes showed a similar response in all growth conditions indicating their co-transcription. Their microarray work showed 13 CDSs that had similar expression profiles at two or three CO_2 concentrations. These included genes involved in amino acid metabolism, production of antioxidant thioredoxins and a CDS potentially implicated in biofilm formation. When the authors examined the transcriptional profile of clostripain, the putative toxin activator, they found that it appeared to be constitutively expressed. It is therefore likely that clostripain contributes to general proteolytic activity in addition to its assumed role as toxin activator. One area where there were significant effects with increasing CO_2 concentration was in spore germinant receptors and spore coat protein expression. If CO_2 concentrations influenced the resistance of spores then this would be of great interest to the food industry where CO_2 is used as an antimicrobial gas.

1.6. Research aims

C. botulinum produces one of the deadliest toxins known to man, the botulinum neurotoxin, which is encoded alongside the associated, non-toxic proteins (ANTPs). The overall objective of this project is to investigate the importance of the ANTPs to *C. botulinum*.

This objective will be addressed via a series of aims:

1. *In silico* analysis of the toxin complex protein sequences with a focus on the identification of related proteins.
2. Investigation of the whole supernatant proteome of *C. botulinum* with a focus on characterisation of potential novel virulence factors.
3. Characterise the toxin complex components produced by a range of clinical isolates.
4. Explore the association between toxin complex type, genomic similarity and disease type.
5. Investigate the transcriptome of *C. botulinum* across a time course to give insight into the transcription of *bont* and the wider botulinum transcriptome.

2. Methods

All chemicals were supplied by Sigma Aldrich (Gillingham, UK) unless otherwise stated. All molecular biology reagents were supplied by Invitrogen (Paisley, UK) unless otherwise stated.

2.1. In silico investigation of the botulinum toxin complex and C. botulinum proteome

2.1.1. In silico investigation of botulinum neurotoxin and the neurotoxin associated proteins

The amino acid sequences of BoNT, the ANTPs and BotR were analysed using BLASTp via the NCBI web server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). To aid communication of the results of the BLASTp analysis a simple similarity score was used. This was the coverage multiplied by the identity, for example, if the matched protein showed 30% identity across 96% of the query protein, a similarity score of 0.29 was calculated. In some scenarios, an alternative protein similarity detection tool, HMMer (http://hmmer.janelia.org/) was used. Typically, this was to confirm or repudiate a low similarity match. Interproscan (Quevillon et al., 2005) is a tool that combines different protein signature recognition methods in one web server (http://www.ebi.ac.uk/Tools/pfa/iprscan/), it was used to identify conserved protein domains in amino acid sequences of interest.

To obtain dendrograms for the protein sequences a distance matrix of the amino acid sequences was determined using MEGA (Tamura et al., 2011) and a dendrogram derived from the distance matrix using a Fitch-Margoliash least-

squares method for clustering based on genetic distance implement in PHYLIP v3.69 (Felsenstein, 1989) and was visualised in Dendroscope (http://ab.inf.uni-tuebingen.de/software/dendroscope/). The 16S phylogeny of the species encoding P-47 family gene clusters was determined using a Maximum Likelihood tree implemented in MEGA.

The genomic environment of the P-47 family proteins was investigated using Artemis (http://www.sanger.ac.uk/resources/software/artemis/) to load and visualise annotated genomes (Genbank format) of the relevant species (Rutherford et al., 2000).

2.1.2. Protein sub cellular location

The publically available genome of *C. botulinum* A1 ATCC 19397 was analysed using tools which predict the sub cellular localisation of proteins. The coding sequences (CDSs) were analysed using the web interface for PsortB (Yu et al., 2010), CELLO (Yu et al., 2006), LocateP (Zhou et al., 2008), , SecretomeP (Bendtsen et al., 2005) and SignalP (Bendtsen et al., 2004). The results of this analysis were then compared using pivot tables in Excel and a consensus formed.

- PsortB is specifically designed for bacteria, discriminating between 4 possible subcellular localisations for Gram-positive bacteria (cytoplasm, membrane, cell wall and extracellular). It uses a combination of BLAST homology to proteins of known subcellular localisation, PROSITE motifs and profiles, signal peptides and trans-membrane helix predictors based on HMMs and one Support Vector Machine (SVM, a supervised learning technique involving training on a known dataset) based prediction module for each localisation using the occurrence of frequent sub-sequences. It has been designed with precision in mind and emphasises specificity over

accuracy. PsortB v 3.0 improves the coverage of the proteome, assigning more proteins a subcellular location while maintaining precision. The high precision means many proteins (~20%) aren't assigned a subcellular location.

- CELLO 2.5 is designed to work with eukaryotic and Gram negative or positive bacteria. It uses a multi-layered SVM approach with the first layer making predictions and the second layer combing those predictions and providing a final assignment. Four types of sequence coding schemes are taken into account: the amino acid composition, the di-peptide composition, the partitioned amino acid composition and the sequence composition based on the physico-chemical properties of amino acids. The outputs from these classifiers are then combined and a final assignment made. CELLO has a higher accuracy compared to PsortB but a lower precision.
- LocateP combines many of the existing high-precision subcellular location identifiers. It provides a higher degree of specificity than the other tools tested distinguishing 7 different subcellular locations in Gram positive bacteria - intracellular, multi-transmembrane, N-terminally membrane anchored, C-terminally membrane anchored and secreted/released proteins. It also distinguishes between Sec- and Tat- It has a high degree of accuracy, especially for Gram positive bacteria, and a high level of detail with the ability to select either the LocateP prediction or the prediction by SwissProt classification.
- SecreteomeP uses a neural network trained on a set of proteins that are known to be secreted by non-classical (i.e. non-signal peptide) secretion pathways.

- SignalP consists of two predictors based on neural networks and hidden Markov model algorithms. It analyses a protein sequence for the presence of a signal peptide.

For PSORTb v3 the data were downloaded from a database of pre-computed genomes, any protein for which the primary assignment was extracellular was counted as such. For analysis in CELLO each genome the coding sequences was downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) and used as the input for a CELLO search. Any protein for which the only assignment was extracellular was counted as such. CELLO classifies many proteins into more than one subcellular localisation but since it is known that CELLO already favours accuracy over specificity (i.e. it returns more false positives) it was decided to narrow down the categorisation to proteins that were assigned a single localisation. LocateP can output pre-computed results for a range of bacterial genomes. As mentioned above the output can be either the LocateP prediction or a prediction bassed on the SwissProt classification. The LocateP algorithm predicts a subset of the SwissProt predictions with higher certainty. The proteins which the SwissProt based algorithm predicted as extracellular were used as the LocateP algorithm gives a small number of results. However, it will be interesting to see whether the higher certainty associated with the LocateP predictions gives better agreement with the other tools.

2.2. Microbiology

2.2.1. Strain list

Table 6: Strains used in this study, WB = wound botulism, FB = food botulism, IB = infant botulism, FPRU – Foodborne Pathogen Reference Unit at HPA

Species	Strain	Toxin type	Isolated from	Source of strain
C. botulinum	ATCC 19397	A1	Type strain	Type strain - FPRU collection
C. botulinum	NCTC 2012	A3	Type strain - food botulism	NCTC
C. sporogenes	NCTC 275	N/A	Gas gangrene	NCTC
C. botulinum	NCTC 2916	A1(B)	Type strain	NCTC
C. butyricum	NCTC 7423	Non-toxigenic	Type strain	NCTC
C. botulinum	H040660361	A	WB	Clinical - isolated by FPRU
C. botulinum	H040680341	A5(B)	WB	Clinical - isolated by FPRU
C. botulinum	H042440055	A5(B)	WB	Clinical - isolated by FPRU
C. botulinum	H044020065	A5(B)	WB	Clinical - isolated by FPRU
C. botulinum	H044640107	A5(B)	WB	Clinical - isolated by FPRU
C. botulinum	H052880114	A	WB	Clinical - isolated by FPRU
C. botulinum	H062260493	B	WB	Clinical - isolated by FPRU
C. botulinum	H063740588	AB	FB	Clinical - isolated by FPRU
C. botulinum	H063960325	A	WB	Clinical - isolated by FPRU
C. botulinum	H064620409	A	WB	Clinical - isolated by FPRU
C. botulinum	H065060505	A	WB	Clinical - isolated by FPRU
C. botulinum	H065260139	A	WB	Clinical - isolated by FPRU
C. botulinum	H071040476	A	WB	Clinical - isolated by FPRU
C. botulinum	H071400014	A	WB	Clinical - isolated by FPRU
C. botulinum	H074240407	A	IB	Clinical - isolated by FPRU
C. botulinum	H074400585	B	IB	Clinical - isolated by FPRU
C. botulinum	H074400586	B	IB	Clinical - isolated by FPRU
C. botulinum	H075000578	A	WB	Clinical - isolated by FPRU
C. botulinum	H090840606	B	IB	Clinical - isolated by FPRU
C. botulinum	H091140481	B	WB	Clinical - isolated by FPRU
C. botulinum	H091280045	B	WB	Clinical - isolated by FPRU
C. botulinum	H091640054	AB	WB	Clinical - isolated by FPRU
C. botulinum	H092080005	B	WB	Clinical - isolated by FPRU
C. botulinum	H093320637	A	IB	Clinical - isolated by FPRU
C. botulinum	H093620104	B	WB	Clinical - isolated by FPRU
C. botulinum	H094460264	A	IB	Clinical - isolated by FPRU
C. butyricum	H102020560	E	IB	Clinical - isolated by FPRU
C. butyricum	H102020561	E	IB	Clinical - isolated by FPRU
C. botulinum	H102120680	B	WB	Clinical - isolated by FPRU
C. botulinum	H10212680	B	WB	Clinical - isolated by FPRU
C. butyricum	H110141052	E	IB	Clinical - isolated by FPRU
C. butyricum	H110220838	E	IB (adult)	Clinical - isolated by FPRU
C. butyricum	H110340631	E	IB (environment)	Clinical - isolated by FPRU
C. butyricum	H110480771	E	IB	Clinical - isolated by FPRU
C. butyricum	H110660554	E	IB	Clinical - isolated by FPRU
C. botulinum	H111860974	A	IB	Clinical - isolated by FPRU
C. botulinum	H112480657	A	IB	Clinical - isolated by FPRU

C. botulinum	H113660204	A	WB	Clinical - isolated by FPRU
C. botulinum	H114400598	B	IB	Clinical - isolated by FPRU
C. botulinum	H114580650	A	FB	Clinical - isolated by FPRU
C. botulinum	H114580654	A	FB	Clinical - isolated by FPRU
C. botulinum	H114590007	A	FB - environmental	Clinical - isolated by FPRU
C. botulinum	H114680633	A	FB	Clinical - isolated by FPRU

2.2.2. Growth conditions

Cells were maintained either on beads at -80°C or in Cooked Meat Broth (CMB) and then revived by anaerobic incubation (10% CO_2, 5% H_2, 85% N_2) on *Clostridium botulinum* isolation agar (CBI) (Oxoid, UK) at 35°C for 48 h to obtain luxuriant growth and ensure pure culture. Starter cultures in either 20 or 100 ml of Tryptone Peptone Glucose Yeast extract (TPGY) (Oxoid, UK) broth were inoculated with single colonies from CBI plates and incubated for 24 h at 35°C. The starter cultures were then used to inoculate either 20 ml of TPGY to an optical density (OD) measured at 600 nm of 0.15-0.2. All media were pre-reduced for 24 h under anaerobic conditions before use. These growth conditions were used as they are the accredited best practice employed by the Foodborne Pathogens Reference Laboratory at PHE Colindale. TPGY was used for culture as it is widely used in other similar studies (Cheng et al., 2008; Bradshaw et al., 2004).

2.2.3. Growth curves

In order to determine growth curves for *C. botulinum* and *C. sporogenes* 100 ml of pre-reduced TPGY was inoculated to an OD600 of 0.15-0.2 from an overnight broth culture of *C. botulinum*. Samples (1 ml) were then taken at appropriate intervals for the period of time being studied. This was typically 12-16 h when RNA work was being carried out or 24 or 96 h if protein work was being performed.

Samples were then analysed at 600 nm using the Eppendorf BioPhotometer spectrophotometer and the optical density recorded. If the total cell count was being carried out then the sample was fixed with 0.1 volume 10% formalin, stained with 1-2 µl crystal violet and then diluted to an appropriate factor to allow for manual counting using a disposable haemocytometer (C-Chip, Lab Tech International, East Sussex) and a microscope. The resulting count was then multiplied by the dilution factor to give number of cells per ml of broth.

2.3. Investigation of the proteome of C. botulinum

2.3.1. Protein precipitation

Two protein precipitation methods were used for this work; tricholoroacetic acid (TCA) precipitation and acetone precipitation. The protocol for TCA precipitation was as follows. A volume of culture supernatant was taken, centrifuged at 12000 g, 4°C for 10 minutes to pellet the cells, filtered through 0.22 µm filter (Millipore, Watford, UK) to ensure it was cell free. For acid precipitation 3 N TCA was added to 5%, 10% or 20% (v/v) and the supernatant left on ice for 30 mins, pelleted by centrifugation at 12000 g, 4°C for 10 minutes, washed with one volume of 96% ethanol followed by one volume of acetone and resuspended in 50mM Tris-HCl, 1% SDS (Gibco, Invitrogen, Paisley, UK) (v/v). For acetone precipitation, 3 volumes of acetone was added to 1 volume of cell free culture supernatant and left on ice for 30 mins, washed with one volume of 96% ethanol followed by one volume of acetone and resuspended in 50mM Tris-HCl, 1% SDS (v/v) (Gibco, Invitrogen, Paisley, UK).

2.3.2. Determination of protein concentration

Total supernatant protein concentration was determined using the Bradford colourimetric protein assay (Bradford, 1976). Bovine Serum Albumin protein standards of 0, 0.05, 0.125, 0.2, 0.5, 0.75, 1, 1.5 and 2 mg/ml were made up in TPGY broth. To obtain the standard curve 5 µl of each standard was added to a flat bottomed 96 well microtitre plate (Sterilin, ThermoFisher, Newport, UK) in triplicate. Then 5 µl of each sample being quantified was added to the microtitre plate in two different dilutions (typically undiluted and 1 in 5 dilution). Following this 250 µl of Bradford reagent was added, the plate sealed and shaken and incubated at RT for 10 mins and then read using a microplate reader (ELx808, Biotek, Bedfordshire, UK) at 595 nm.

The concentration of precipitated proteins which had been resuspended in 50 mM Tris-HCl, 1% SDS (v/v) was determined using the detergent compatible bicinchoninic acid (BCA) assay (Pierce, Thermo Scientific, Northumberland, UK). Protein standards of 0, 0.05, 0.125, 0.2, 0.5, 0.75, 1, 1.5 and 2 mg/ml were made up in 50 mM Tris-HCl, 1% SDS. Then 25 µl of each standard and each sample being quantified was added to a microtitre plate in triplicate. Two dilutions, typically 1 in 5 and 1 in 10 dilution of the unknown sample, were typically assayed. Subsequently, 250 µl of BCA reagent working solution was added, the plate sealed, briefly vortexed and incubated at 37°C for 30 mins. After cooling to RT the plate was briefly centrifuged and read at 565 nm. The gradient of the standard curve was calculated and used to convert the absorbance value of the samples into protein concentrations.

2.3.3. Endopeptidase immunoassay for measuring botulinum toxin activity

BoNT activity in the culture supernatant was assayed using an in vitro endopeptidase activity immunoassay according to the method of Jones et al, 2008. A polystyrene 96-well plate (Nunc Maxisorp) were coated with 2 µg/ml $SNAP25_{137-206}$ peptide substrate in 50 mM carbonate buffer (pH 9,6) at 4°C and then washed 3 time with PBST (phosphate buffered saline with 0.5%mv/v Tween 20) and blocked with 300 µl/well of 5% skimmed milk powder in PBST for 90 min at RT. To test the concentration of toxin present in the sample, 100 µl of 1 in 100 diluted culture supernatant was added to the plate in duplicate. Each duplicate was serially two-fold diluted 8 times and incubated overnight at RT. A positive control of a two fold dilution series of a BoNT standard of known concentration was included on each plate. After incubation, the sample was removed and the plate was washed 3 times with PBST. After washing, 100 µl of primary detecting antibody, specific to the cleaved substrate epitope (anti-$SNAP25_{(190-197)}$ was added and incubated for 90 mins at RT. The plate was then washed again and 100 µl/well of goat anti-rabbit-HRP conjugate (Sigma A0545, 1 in 2000 dilution in antibody buffer) was added and incubated for 90 mins at RT. After washing, 100 µl/well of HRP-substrate solution (50mMcitric acid pH 4.0, 0.05% w/v ABTS (2,2′-Azino-bis(3-ethylben- zothiazoline-6-sulfonic acid) diammonium salt, and 0.05% v/v of a 30% w/v Hydrogen peroxide solution) was added and colour allowed to develop at RT for 30 min. The absorbance was then read at 405 nm using a Multiscan plate reader. The relative activity of botulinum toxin in the culture supernatant was calculated in reference to the known standard.

2.3.4. 1D Sodium Dodecyl Sulphate Polyacrylamide Gel electrophoresis (1D SDS-PAGE)

Briefly, 5 µg of protein sample was combined with 1x NuPAGE LDS Sample Buffer, 1x NuPAGE Reducing Agent made up to 10µl with deionised water and heated at 70°C for 10 minutes. The sample was loaded onto the gel alongside 5 µl Mark12 Unstained Standard (Invitrogen). Typically both 12% acrylamide bis-tris gels and 7% Tris-Acetate gels were run in order to obtain high quality separation across the majority of molecular weights relevant to this project. Gels were run using the NuPAGE system (Invitrogen, Paisley UK) and either MOPS buffer (for the bis-tris gels) or TA buffer (for the Tris-Acetate gels). After running, the gels were stained using Instant Blue (Expedeon, Cambridgeshire, UK) and imaged using the ProPic II spot picker (Digilab, Cambridgeshire, UK).

For analysis the gel image was loaded into the ImageQuant TL software (v. 2005, Amersham Biosciences, GE Healthcare, Buckinghamshire, UK) which automatically assigned lanes and performed background subtraction. Bands were then manually picked in each lane, the size calculated by assigning the ladder bands their known weights and the band volumes normalised according to the 200 kDa ladder band allowing for comparison between gels.

2.3.5. In-gel digestion

When samples were to be analysed by LC-MS/MS the appropriate 1D gel lane was divided into 12 sections. The sections were destained by washing with 50% 50mM ammonium bicarbonate/50% methanol. Following dehydration with acetonitrile the plugs were incubated with 50 µl 10 µM dithiothreitol (DTT) followed by 10 µl 55 mM iodoacetamide (IAA) to ensure reduction and alkylation of the

proteins. After washing with 50 mM ammonium bicarbonate and dehydration with acetonitrile the gel plugs were then incubated for 17 h with 50 µl of 10 ng/µl mass spectrometry grade trypsin (Promega, Southampton, UK). Following trypsinisation 50 µl of 0.1% trifluoroacetic acid was added to the gel plug and incubated at RT for 1 h and the supernatant transferred to a clean microtitre plate and stored at -80 °C until analysed.

2.3.6. LC-MS analysis of peptide fragment mixtures

The peptides resulting from in gel digestion were separated by liquid chromatography (LC) and analysed by mass spectrometry. Running on the liquid chromatography column separates the peptides by hydrophobicity. Subsequently, an MS/MS approach was used whereby the peptides are ionised in MS mode and specific ions are fragmented further in MS/MS mode. The resulting spectra can be analysed using appropriate software which compare the experimental results against an *in silico* database of peptide fragments. The peptide fragments which the algorithm identifies as being present in the sample can then be matched to their 'parent' proteins and, if the analysis criteria are satisfied then the protein is 'identified' as being present in the sample.

The system used for this was an Ultimate 3000 Dionex nano/capillary HPLC system (Dionex) coupled to a LTQ Orbitrap mass spectrometer (Thermo Electron). The data produced was in the Thermo Finnigan .RAW file format. Mascot (Matrix Science, UK) was used to analyse the .RAW files, the parameters were Enzyme = Trypsin; fixed modifications = carbamidomethylation of cystine; Variable Modifications = oxidation of methionine; Missed Cleavage Sites: 2; peptide mass tolerance ±10 ppm. The data from this analysis was visualised using Scaffold (Proteome Software, USA).

The mobile phase was, buffer A; 2% acetonitrile in water plus 0.1% formic acid, buffer B; 10% water in acetonitrile plus 0.1% formic acid. Injection volume was 10 µl, flow rate loading pump 25 µl/min, micro pump: 300 nl/min. The gradient used was:

- Start 0 min: 10% B
- 0 – 40 min: 10% B – 45% B
- 40 – 40.2 min: 90% B
- 40.2 – 52 min: 90% B
- 52-52.2 min: 10% B
- 52.2 – 60 min: 10% B

The mass spectrometer was operated in positive mode with spray voltage at 1.6 kV, capillary voltage at 38 V, capillary temperature at 200°C and tube lens at 125 V. Helium was used as collision gas but no sheath and auxiliary gas were applied. Tandem MS (MS/MS) data was acquired in data-dependent mode with automatic switching between MS and MS/MS modes. A normalised collision energy of 35%, an activation of q = 0.25 and activation time of 30 msec were applied in MS/MS acquisition. The precursor ion scan (m/z 440-2000) were acquired in the Orbitrap with a resolution R = 60000 at m/z 400. The six most abundant peptide precursor ions detected in the preceding survey scan were dynamically selected and subjected for collision-induced dissociation (CID) in the linear ion trap to generate MS/MS spectra. The lock mass option, using the polydimethylcyclosiloxane ion generated in the electrospray process from ambient air, the protonated (Si(CH3)2O)6 at m/z 445.120025, was used for internal recalibration in real time to enable accurate mass measurement. Samples were analysed in a format of technical triplicates. Software used was Xcalibur 2.1.0 sp1.1160.

The Thermo RAW files were loaded directly into Mascot Daemon v2.2.2 using the input filter *'ThermoFinnigan LCQ / DECA RAW file'*. The databases searched depended on the analysis being performed but were either a non-redundant version of the NCBI protein database or an orgamism specific database using the .faa files from Genbank. Mascot parameters were; fragment tolerance 0.1 Da, parent tolerance 10.0 parts per million (ppm), digestion enzyme trypsin and max missed cleavages 2. Scaffold (version Scaffold_3.6.5, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 50.0% probability as specified by the Peptide Prophet algorithm (Keller et al., 2002 Anal. Chem. 2002;74(20):5383-92). Peptide identifications were also required to exceed specific database search engine thresholds. Mascot identifications required at least ion scores must be greater than both the associated identity scores and, 20, 30 for, doubly, triply charged peptides. Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii, 2003). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

2.3.7. Comparison with MvirDB

The proteins which were detected and identified in the *C. botulinum* supernatant were compared with the virulence database MvirDB using local BLAST (Zhou et al., 2007). The homologous protein pairs were parsed for potential virulence factors which could play a role in food, wound or infant botulism. Proteins which had a putative role were then analysed in greater depth to ensure that the areas of

homology were not generic regions with no pathogenic connotations e.g. transmembrane regions. This was done using Interproscan (SwissProt) to determine the location of different functional groups. If the *C. botulinum* extracellular protein showed homology, relevant to virulence functional groups with the virulence associated protein from the database then it was considered a putative virulence protein.

2.3.8. Calculation of extracellular protein cost

The number of high energy phosphate bonds required to synthesise each amino acid in chemoheterotrophs was obtained from the literature (Heizer et al., 2006). The amino acid content of each identified extracellular protein was obtained from NCBI and the cost of synthesis of each protein calculated using a custom Perl script.

2.4. Investigation of toxin complex gene type and genomic background of botulinum toxin producing strains

2.4.1. DNA extraction

DNA was extracted using either MicroLYSIS (Microzone, UK) reagent or DNeasy columns (Qiagen). For MicroLYSIS a 1 µl loop was touched to a single bacterial colony and then resuspended in 19 µl of MicroLYSIS reagent in a PCR tube. This tube was then placed in a PCR machine with the following programme:

Step 1: 65°C for 5 mins
Step 2: 96°C for 2 mins

Step 3: 65°C for 4 mins
Step 4: 96°C for 1 mins
Step 5: 65°C for 1 mins
Step 6: 96°C for 30 secs
Step 7: 20 °C hold

Once the programme was finished the sample was either placed at 4°C if being processed the same day or at -20°C if being processed at a later date.

For extraction using DNeasy columns a 10 µl loop of bacterial growth was resuspended in 200 µl Phosphate Buffered Saline and 20 µl proteinase K (18 mg/ml) (Roche, Burgess Hill, UK). Then 200 µl DNeasy Buffer AL was added, the sample vortexed and incubated at 56°C for 10 minutes. The mixture was then pipetted into a spin column and centrifuged for 1 min at 12000 rpm. The column was then washed with 500 µl wash buffer AW1 and centrifuged for 1 min at 12000 rpm followed by 500 µl wash buffer AW2 and the column centrifuged for 3 min at 12000 rpm. The column was then transferred to a new tube and 200 µl elution Buffer AE added, incubated for 1 min and then centrifuged for 1 min at 12000 rpm. Extracted DNA was stored at 4° C if used with two weeks or at -80°C for long term storage.

2.4.2. PCR to determine toxin type and toxin complex type

Two types of PCR were used in this investigation, block -based PCR and qPCR. qPCR assays designed for use in the reference lab (Akbulut et al,. 2004) were used to detect *bont/A* and *bont/B*. Standard PCR was used to determine the toxin complex type using an assay designed for this study comprising two pairs of primers specific to *ha70* and *ha33* and two pairs of primers specific to *orfX/2* and *orfX/3*. The primers used in this investigation can be seen in Table 7.

2.4.3. Fluorescent amplified fragment length polymorphism (fAFLP) analysis of clinical strains

Genomic DNA was digested using restriction enzymes Hind III and Hha I and the resulting fragments ligated to the appropriate oligonucleotide adaptors (Table 7) in a one step reaction. Adapters were annealed prior to this by heating complimentary strands to 94°C for 10 minutes and allowing to cool to room temperature. Digestion and ligation conditions consisted of 5 µl of DNA, 3 µl of 10x NEB buffer 2, 0.3 µl of 100x BSA (NEB), 5 µl of 10x T4 Ligase buffer (NEB), 0.2 µl of 400 U/µl T4 ligase (NEB), 0.3 µl of *HindIII* 20 U/µl restriction endonuclease (NEB), 0.3 µl of 20 U/µl *HhaI* restriction endonuclease (NEB), 0.3 µl of 2 µM annealed *HindIII* adaptors (table 2) (Eurogentec), 0.3 µl of 20 µM annealed *HhaI* adaptors (table 2) (Eurogentec), 0.5 µl of 30 mg/ml RNAse (Sigma) and 44 µl of molecular grade water. The reaction mixture was incubated at 37°C for 4 hours, 16°C for 3 hours then 65°C for 10 minutes. The resulting fragments were then amplified using a touch down PCR reaction comprised of 5 µl of ligated fragments, 12.5 µl 2x MegaMix Gold PCR mastermix (Microzone), 2.5 µl 10 µM *Hind-0* primer (table 2) and 2.5 µl 10 µM Hha-A primer (table 2). The reaction mixture was then incubated at 95°C for 5 mins, 10 cycles of 94 °C for 20s, 66°C for 30s (reduced by 1°C per cycle) and 72°C for 2 mins followed by 20 cycles with the annealing temperature at 56°C. The Hha I primer included the selective 3' A base while the Hind III primer had a 6'-carboxyfluorescein label incorporated and no selective bases. From the resulting product, 1 ul was added to 10 µl Hi-Di formamide and 0.5 µl LIZ 600 labelled standard, denatured at 95°C for 5 minutes and cooled to 4°C. Samples were loaded on to a 3730 ABI sequencer in accordance with the manufacturer's guidelines by a local sequencing service. Reproducibility was

assessed by carrying out the entire fAFLP procedure (digestion-ligation-PCR-fragment analysis) procedure in triplicate for 9 samples; this allowed us to distinguish between biological and technical variation in the procedure. When these samples were clustered the average similarity between each group of three replicates was 94.9% with a range of 91.4-96.9%. No differentiation was inferred when the similarity between two strains was greater than 90% as it cannot be determined if the difference is the result of technical or biological variation.

Electropherograms were analysed using Bionumerics v6.1 (Applied Maths, Belgium). Fragments sized between 60 and 600bp were used for further investigation. Profiles were manually inspected using Bionumerics and Peakscanner (Applied Biosystems, USA). Pearsons correlation coefficient for the samples was calculated and the samples clustered using UPGMA, both performed within Bionumerics.

2.5. Analysis of transcription in C. botulinum

All work involving RNA was carried out in an area cleaned by RNase ZAP (Ambion, Life Technologies, Paisley, UK) using machinery cleaned with RNase ZAP in order to reduce RNase contamination. All tubes and tips were RNase free (Ambion, Paisley, UK) and all water used was DEPC treated (Severn Biotech, Worcester, UK).

2.5.1. Extraction of RNA from C. botulinum culture

To prevent RNA degradation one volume (typically 500 µl) of culture was removed and added to 2 volumes of RNAprotect Bacteria (Qiagen), vortexed for 5 s and incubated at RT for 5 min. Samples were then centrifuged for 10 minutes at 5000 g, the supernatant discarded and the sample either lysed immediately or stored at -80°C.

Two different lysis methods were investigated in order to determine the methodology that gave optimal RNA yield and quality; enzyme lysis and bead beating lysis.

For enzyme lysis 200 µl of TE buffer (10 mM Tris-Cl, 1 mM EDTA, pH 8.0) containing 1 mg/ml lysozyme was added to the cell pellet, the sample was vortexed for 10 s and incubated at room temperature for 10 min with mixing every 2 min. Then 700 µl RLT buffer from the RNeasy kit with β-meracaptoethanol was added, the sample was vortexed, centrifuged and the supernatant added to a new tube. A 500 µl volume of 100% ethanol was added to the sample before purification of the RNA using the RNeasy mini kit (Qiagen) with on column DNase I (Invitrogen) treatment according to the manufacturer's instructions.

For bead beating lysis the bacterial sample was resuspended in 700 µl buffer RLT from the RNeasy Mini kit (Qiagen, West Sussex, UK) with added β-meracaptoethanol and approximately 50 mg glass beads were added. The cells were disrupted in a bead beater for 30 s, 3 times with a brief rest in between and then left on ice for 1 min. The tubes were centrifuged at 21000 g for 1 min and the supernatant transferred to a new tube. The volume of the supernatant was determined and if column extraction was being performed then an equal volume of 70% ethanol was added before RNA purification using the RNeasy mini kit

(Qiagen) with on column DNase I (Invitrogen) treatment according to the manufacturer's instructions. If 25:24:1 extraction was being performed then the methodology outlined in 2.5.9 was carried out.

The RNeasy mini kit does not capture small RNA (sRNA, <200 bp). In order to purify these sRNAs the flow through of the RNeasy columns was captured and the RNA purified from it using 25:24:1 purification detailed in 2.5.9.

2.5.2. DNase treatment

Genomic DNA has to be removed before reverse transcription qPCR in order to allow accurate assessment of transcript abundance. The amount of gDNA was reduced below levels detectable by qPCR for *bont/A* and *gdhA* using TURBO DNA-*free* (Ambion, Life Technologies, Paisley, UK). Briefly 0.1 volume 10x Turbo DNase buffer and 1-5 µl Turbo DNase were added to the RNA extract, mixed gently and then incubated at 37°C for 30 mins. After centrifugation to collect condensation, 0.1-0.2 volumes of TURBO DNA-*free* DNase inactivation reagent were added, the sample mixed and incubated at room temperature for 5 mins with occasional mixing before centrifugation at 10 000 x g for 90 s. The supernatant was then transferred to a new tube and the presence of contaminating gDNA assayed using qPCR assays for *bont/A* and the house keeping gene *gluD*. If gDNA could still be detected the sample was purified using phenol:chloroform:isoamyl alcohol purification and DNase treatment repeated until gDNA could no longer be detected by qPCR targeting *bont* and *gdhA*

2.5.3. RNA quantity and purity analysis

The RNA content of the DNA-free RNA extract was estimated by UV spectrophotometry using the NanoDrop 8000. This gives RNA concentrations and 260/280 and 260/230 absorption values. A 260/280 ratio of greater than 2.0 was taken as 'pure' RNA while low 260/230 absorption values (<1.5) indicated contamination with carbohydrates or phenol.

When more precise RNA quantification was required the Quant-iT Ribogreen kit (Invitrogen, Paisley, UK) was used according to the manufacturer's instructions.

2.5.4. Analysis of RNA quality by agarose gel electrophoresis

Agarose gels were made up using 1% (w/v) agarose in 1x TAE buffer, GelRed safe stain was added to the agarose at 1x concentration. Gels were run at 80-120 V in 1x TAE buffer. Prior to loading 10 µl of sample was combined with 1.7 µl 6x loading buffer. A 100 bp DNA ladder was used for determining size. Gels were visualised using the BioRad imager.

2.5.5. Analysis of RNA quality using Agilent Bioanalyser

RNA was analysed using the Bioanalyser RNA 6000 Nano kit. A gel matrix-dye solution was made up using 1 µl of dye added to 65 µl filtered gel matrix which was mixed and centrifuged. Then 9 µl of gel-dye mix was placed in the primary gel-dye mix well on an RNA 6000 Nano chip which was placed in a chip priming station. The syringe plunger on the chip priming station was positioned at 1 ml and the chip priming station closed. The plunger was depressed until held by the clip and after exactly 30 seconds the clip was released. After 5 s the plunger was

pulled back to 1 ml and the chip priming station opened, 9 µl of gel-dye mix was added to the secondary gel-dye mix wells and 5 µl of marker was added to every sample well and the ladder well. A 1 µl aliquot of RNA Nano 6000 ladder was added to the ladder well and 1 µl of sample was added to each sample well with any unused wells having 1 µl of marker added. The chip was vortexed in an IKA vortexer at 2400 rpm for 1 min and then run on the Agilent 2100 bioanalyser within 5 minutes.

2.5.6. Reverse Transcription

Reverse transcription was carried out using SuperScript III Reverse Transcriptase (Invitrogen, Paisley, UK) in a 20 µl reaction with 0.5 µg total RNA, 2 pmol of gene specific, reverse primer for *bont* and *gluD*, 1 µl of 10 mM dNTP mix (invitrogen, Paisley, UK). First the primers, RNA and dNTPs were combined and made up to 13 µl with water before being heated to 65°C for 5 mins, cooled on ice and briefly centrifuged. Then 4 µl 5x First-Strand Buffer, 0.1 µl 0.1 M DTT, 1 µl RNaseOUT (Invitrogen, Paisley, UK) and 1 µl SuperScript III reverse transcriptase added to each reaction. This reaction was then placed in a thermal cycler heated to 25°C for 5 minutes, 55°C for 45 minutes and 70°C for 15 minutes. The reverse transcription for *bont/A* and *gluD* were carried out separately. Resultant cDNA was quantified using the Nanodrop and 1 µg cDNA used for subsequent qPCR

2.5.7. Quantitatiive PCR (qPCR)

The Taqman 7500 Fast Real Time PCR system was used for quantitative PCR analysis. The standard PCR mix used was a 25 µl reaction consisting of 0.5 µg RNA, 12.5 µl TaqMan Fast Universal PCR Master Mix (2x), 0.75 µl 10 µM primers

(Eurogentec, Hampshire, UK), 0.5 µl 5 mM probe (Eurogentec, Hampshire, UK) and target sample and SDW to 9.5 µl. To ensure reproducible determination of the crossing threshold (Ct) values, it was the point at which the sample in the positive control well entered logarithmic amplification.

The efficiency of each qPCR reaction (*bont/A* and *gluD*) was determined by preparing a standard curve using a serially diluted RNA sample of known concentration (from 1 µg/µl to 100 pg/µl). The slope of the standard curve was then used to calculate the RT-qPCR efficiency according to Equation 1.

Equation 1: Calculation of efficiency of the RT-qPCR when the slope of the line of RNA concentration compared with reaction C_t value has been derived.

$$\text{Efficiency} = -1+10^{(-1/\text{slope})}$$

Equation 2: Efficiency dependent model for calculating relative gene expression of target and housekeeping (reference) genes, the control was the earliest time point sampled

$$\text{ratio} = \frac{(E_{\text{Ref}})^{C_P \text{ sample}}}{(E_{\text{target}})^{C_P \text{ sample}}} \div \frac{(E_{\text{Ref}})^{C_P \text{ calibrator}}}{(E_{\text{target}})^{C_P \text{ calibrator}}}$$

2.5.8. Relative gene expression analysis

The relative gene expression of *bont/A* compared to the house keeping gene *gluD* was calculated using the efficiency corrected method (Pfaffl, 2006) detailed in Equation 2.

E_{Ref} is the efficiency of the PCR for the reference gene (*gluD*), E_{target} is the efficiency of the PCR for the target gene (*e.g. bont*), $C_{p\ sample}$ is the crossing point (aka the threshold value or C_t) of the sample, $C_{p\ calibrator}$ is the crossing point of the calibrator sample (in this case the RNA sample from the earliest time point being analysed).

2.5.9. Purification of RNA using phenol:chloroform:isoamyl alcohol

Phenol:chloroform:isoamyl alcohol extraction (phenol pH 4.3, saturated with 0.1 M citrate buffer; chloroform:isoamyl alcohol 24:1) was employed to purify RNA samples containing undesired protein. Two volumes of Phenol:chloroform:isoamyl alcohol in the ratio 25:24:1 was added to one volume of RNA in aqueous solution, thoroughly mixed and centrifuged at 21000 x g for 30 mins at 4 °C. The upper layer was removed, transferred to a new tube and the RNA precipitated by the addition of 3 volumes of 100% EtOH:Sodium acetate (30:1) and incubating at -80 °C for at least 30 mins. The sample was then centrifuged at 21000 x g for 30 mins at 4 °C to pellet the RNA, washed with 70% EtOH, dried using a vacuum drier and resuspended in DEPC-treated water. RNA was stored at -80 °C in aliquots to reduce degradation from freeze-thawing.

2.5.10. Taking of samples for RNA-seq

Once all the methods were optimised the final samples (i.e. two technical replicates of two biological replicates taken at three timepoints; mid-log, late-log and early stationary phase) for RNA-seq were taken. The experimental design allowed two biological replicates to be analysed. The cultures were sampled (3.5 ml) once an hour between mid-log to late stationary phase (approximately 6-13 h after inoculation). Each sample had a 0.5 ml subsample taken from it, both this sample and the 3 ml sample were pelleted and treated with RNAprotect. The 0.5 ml sample from each time point between 6-13 h was analysed to determine the relative gene expression of *bont* compared with *gluD*. The relative gene expression result was used to determine the time point at which toxin expression peaked. Then two time points were chosen, one which was pre-toxin peak and another which was post-toxin peak. The 3 ml sample of the pre-peak, peak and post-peak toxin expression was then processed, the RT-qPCR result confirmed the relative gene expression pattern and the sample was processed for RNA-seq.

2.5.11. Reduction of rRNA before preparation of RNA-seq libraries

rRNA makes up a large proportion (over 90%) of the RNA purified from bacterial cultures. In order to minimise the number of sequencing reads which map to the rRNA genes the rRNA is reduced prior to sequencing.

Each of the 6 samples (6 h A and B, 9 h A, 10 h B, 13 h A and B) were treated with the Ribo-Zero rRNA removal, gram positive bacteria kit (Epicentre, distributed in the UK by Cambio, Cambridge, UK). The Ribo-Zero kit uses rRNA-complementary oligonucleotide probes linked to magnetic beads. The rRNA

probes bind to 5S, 16S and 23S rRNA in the sample, a magnet was then used to pellet the rRNA probes and the rRNA reduced supernatant is removed. The kit were used according to manufacturers instructions.

2.5.12. SOLiD sequencing of samples

The SOLiD RNA-seq protocol was carried out according to manufacturer's instructions by the Eastern Sequencing and Informatics Hub (EASIH), based at Addenbrokes Hospital.

2.5.13. Analysis of RNA-seq data

Gene expression data resulting from RNA-seq is often presented in terms of the reads mapped per kilobase of gene per million mapped reads (RPKM). RPKM is calculated on a gene-by-gene basis according to Equation 3. The advantage of RPKM is that it takes into account both the length of the gene and the total number of reads in a sample. This allows for the comparison of expression level between different genes and between different samples.

Equation 3: Equation to calculate normalised expression level (RPKM)

$$RPKM = \frac{\text{Number of reads which map to gene}}{\text{Total number of reads (Millions)} \times \text{Length of gene (kbp)}}$$

Another way of analysing RNA-seq data is to perform a differential gene expression analysis. This type of analysis typically uses unnormalised expression values (i.e. reads mapped rather than RPKM) to take advantage of the expected negative binomial distribution of gene expression values. The approach used in this work was implemented in edgeR (Robinson et al., 2010), an R package. EdgeR uses an over-dispersed Poisson model to model the count data and an empirical Bayes procedure is used to shrink the dispersions toward a consensus, effectively using gene expression data from all the genes to assess how likely it is that the expression of each particular gene is differentially expressed.

2.6. Primers

Table 7: Primers used in this study

Primer name	Sequence 5'->3'
BoNT/A Fw	CGAAATGGTTATGGCTCTACTCAA
BoNT/A Rev	TTGCCTGCACCTAAAAGAGGAT
BoNT/A Probe	FAM-ACTTCAAGTTGACTCCTCAAAACCAAATGTAAA–TAMRA
GluD Fw	AGTAAAGGTTTTCAAAGGTTATAG
GluD Rev	TCTAAAGAAACATTTGGATGGAA
GluD Probe	NED-CTCAACATAATGATGCAGTAGGTCCAACA-MGB-NFQ
HA70Fw	TAGTGATACTATTGATTTAGCTGATGGT
HA70Rev	CATTTGTTCTTATATATCCATCACCCTT
orfX/2 Fw	CTAATTCCACACTAAATGGTACTTGGAA
orfX/2 Rev	TGATACTGAATTCTCAGTATTCCATCC
ha33 Fw	CAGGATTTATAAGCCAATATTGGGC
ha33 Rev	GCTCGTAATTTGAAGCTTAGCAC
orfX/3 Fw	CCAAGTTATGTAGATTATGCTTATTCAG
orfX/3 Rev	CTATTAAAGCAACTGCTAAAAATGATCC
Hind III 0–primer	5' FAM – GACTGCGTACCAGCTTA – 3'
Hha I A– primer	5' GATGAGTCCTGATCGCA – 3'
Hind III – ST1	5' – CTCGTAGACTGCGTACC – 3'
Hind III – ST2	5' – AGCTGGTACGCAGTC – 3'
Hha I – ST1	5' – GACGATGAGTCCTGATCG – 3'
Hha I – ST2	5' – ATCAGGACTCATCG – 3'

3. Results

3.1. *In silico* investigation of the botulinum toxin complex and *C. botulinum* proteome

Botulinum neurotoxin is encoded and co-expressed within a gene cluster that is 12000 bp to 14000 bp in length, depending on the type of toxin gene cluster subtype. This gene cluster encodes the associated non-toxic proteins (ANTPs). There are two subtypes of toxin gene cluster – the haemagglutinin (HA) and the OrfX.

The HA type gene cluster consists of *bont, ntnH* and *botR* which are encoded on the forward strand and *ha33, ha17* and *ha70* which are encoded on the reverse strand (Figure 11A). In *C. botulinum* A1 ATCC 19397 *bont/A* is 3886 bp long encoding a 149.4 kDa protein, *ntnH* is 3581 bp long encoding a 138.2 kDa protein and *botR* is 536 bp long encoding a 21.7 kDa protein. In ATCC 19397 *ha33* is 881 bp long encoding a 33.8 kDa protein, *ha17* is 440 bp long encoding a 17 kDa protein and *ha70* is 1881 bp long encoding a 71.1 kDa protein.

The genomic arrangement of the OrfX type cluster (Figure 11B) differs from that of the HA cluster. The OrfX cluster consists, with some variants, of *bont, ntnH* and *p47* encoded on the forward strand and *botR, orfX/1, orfX/2* and *orfX/3* encoded on the reverse strand. In *C. botulinum* A3 Loch Maree *bont/A* is 3879 bp encoding a 148.6 kDa protein, *ntnH* is 3480 bp long encoding a 134.6 kDa protein, *p47* is 1248 bp long and encodes a 47.4 kDa protein. On the reverse strand *botR* is 552 bp long and encodes a 22.2 kDa protein, *orfX/1* is 429 bp long and encodes a 16.5

kDa protein, *orfX/2* is 2253 bp long and encodes a 84.3 kDa protein and *orfX/3* is 1473 bp long and encodes a 55.2 kDa protein.

BoNT/A1, A5, B, C, D and G are encoded in the HA-type gene cluster while BoNT/A1 (rarely), A2, A3, A4, E and F have been identified encoded in OrfX gene clusters.

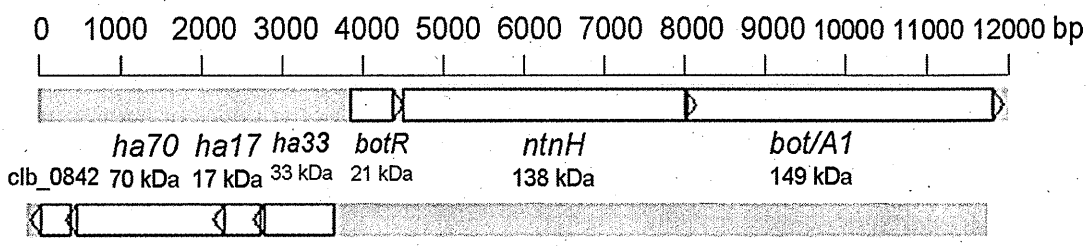
The *bont* and *ntnH* genes are 75-80% similar between the HA and OrfX toxin clusters (Figure 12), the *botR* gene sequences are similar but the gene is in the opposite orientation between the two toxin cluster subtypes, whilst there is no significant similarity between the *ha* genes and the *orfX* genes.

Group I *C. botulinum* cause the majority of botulism in the UK and are the focus of this investigation. Therefore, BoNT, NTNH, BotR and the HA cluster protein sequences from *C. botulinum* A1 ATCC 19397 and the OrfX cluster protein sequences from *C. botulinum* A3 Loch Maree were examined in greater detail.

In this section, BLASTp (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to determine proteins exhibiting significant sequence similarity to the botulinum neurotoxin and the neurotoxin-associated proteins. An E-value cut off of 0.01 was used to determine significance (i.e. only matches with an e-value less than 0.01 were considered significant). For the sake of clarity, identity (percentage of residues which are the same) and coverage (percentage of protein length which exhibits similarity) are multiplied to make a single 'similarity' value. For example, if a protein was 80% similar over 65% of the protein length, 0.8 x 0.65 = 0.52. Identity and coverage values used to derive the similarity values, along with BLASTp E-values are presented in the accompanying tables.

This chapter places the botulinum neurotoxin and associated non-toxin protein sequences into a wider context. It describes the similarity between these proteins in different types and subtypes of BoNT, conserved functional domains within these proteins and their similarity to proteins from other species.

(A)



(B)

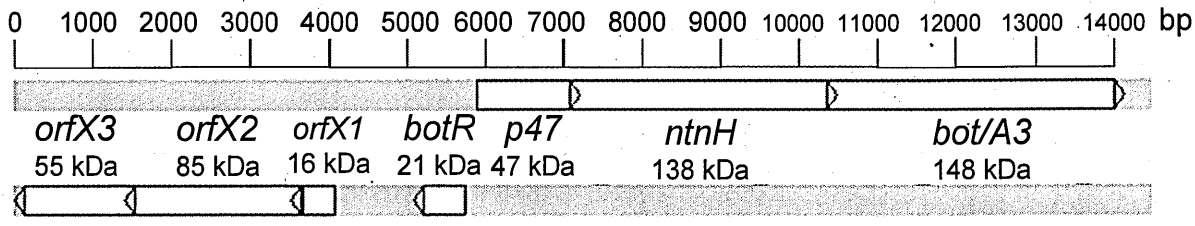


Figure 11: Genomic arrangement of (A) *bont/A1* in a HA gene cluster from *C. botulinum* A1 ATCC 19397 and (B) *bont/A3* from *C. botulinum* A3 NCTC 2012 in an OrfX gene cluster.

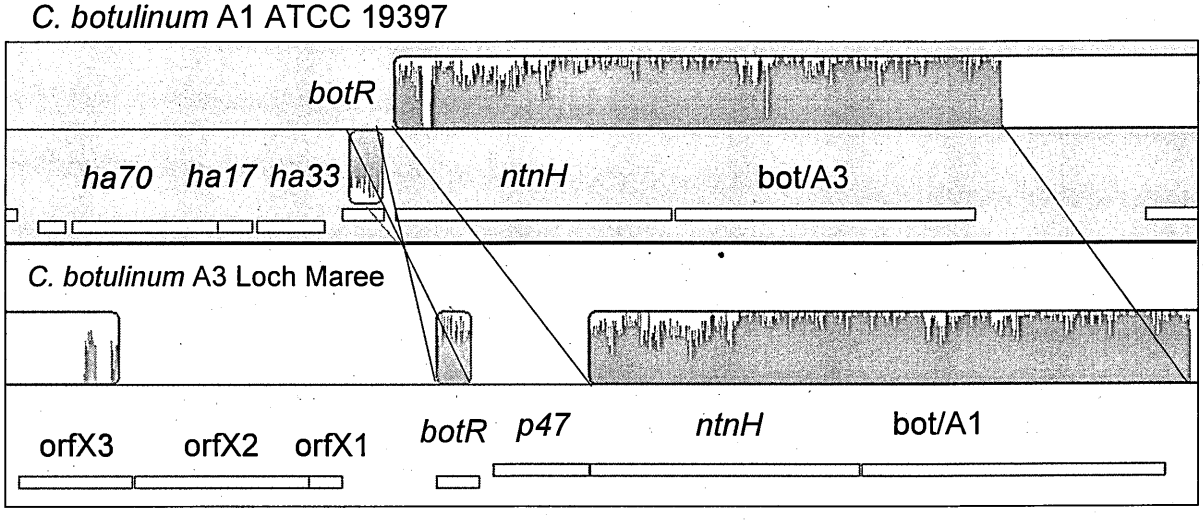


Figure 12: Comparison of *ha* and *orfX* gene clusters. The *C. botulinum* A1 ATCC 19397 *bont/A1* gene encoded within a *ha* gene cluster (above red line) was compared with the *C. botulinum* A3 Loch Maree *orfX* gene cluster (below red line) using Mauve. Areas of similarity are represented with coloured areas, the higher the peaks within the area, the higher the similarity between two areas. The similarity of the *bont, ntnH* and *p47* genes between the two strains can be seen while no similarity between the *ha* genes and the *orfX* genes was detected. The sequence of *botR* is in the opposite orientation between the strains. Created using Mauve v2.3.1.

3.1.1. *In silico* investigation of botulinum neurotoxin and the neurotoxin associated proteins

3.1.1.1. In silico investigation of botulinum neurotoxin A1

The amino acid sequence of BoNT/A from *C. botulinum* A1 ATCC 19397 (Uniprot id = A7FS63) was compared against the NCBI non-redundant protein database using BLASTp. The proteins with the highest similarity to BoNT/A1 were other botulinum neurotoxin proteins. BoNT/A5 has the highest similarity to BoNT/A1 with a similarity score of 0.96 (in this case there was 99% identity over 97% of the protein sequence). BoNT/A2, A4 and A3 have similarity scores of 0.86, 0.84 and 0.79 respectively. BoNT/F, E, G and B have similarity scores of 0.27, 0.27, 0.24 and 0.25 while BoNT/D and C are the least similar with scores of 0.21 and 0.20 respectively. Tetanus neurotoxin (TeNT) is 35.4% similar to BoNT/A1 over 58.9% of the protein length. This gives TeNT a sequence similarity score of 0.21, a comparable level of similarity to the most distant BoNT protein sequences, BoNT/C and D. After TeNT the most similar protein to BoNT/A1 was *C. botulinum* NTNH, which had a similarity score of 0.17.

The protein motifs present in the different BoNT types, TeNT and NTNH were analysed using Interproscan, a program to identify functional domains in protein sequences. All seven BoNT subtypes and TeNT had identical functional domains in an identical arrangement along the length of the protein sequence (Figure 14 A & B). There were four protein domains present in the clostridial neurotoxins; enzymatic domain, translocation domain, receptor binding N terminal and receptor binding C terminal.

The enzymatically active light chain of the toxin (IPR000395) was identified beginning at the N-terminus and finishing, approximately, at amino acid residue 409 (in BoNT/A, and at a very similar position in the other proteins). Clostridium neurotoxin domains responsible for translocation and receptor binding, (IPR012500, IPR012928, IPR013104) were identified in the BoNT/A heavy chain from approximately amino acid residue 526 to the C terminus of the protein (amino acid residue 1296 in BoNT/A). The region responsible for toxin binding to target cells, spans amino acids 760 to 1111 (IPR012928), was also identified as a lectin domain (IPR008985) that may enhance binding to complex carbohydrates such as gangliosides.

A dendrogram (Figure 13) derived from a protein distance matrix concurs with the results of the BLASTp analysis (Table 8) (see methods section X). All the BoNT/A sequences cluster on the same branch, with BoNT/A1 and A5 most closely related. BoNT/C and BoNT/D form a related pair while BoNT/E and BoNT/F form a distinct, related pair although the distance between them is greater than the distance between BoNT/C and BoNT/D. BoNT/B and BoNT/G are on the same branch of the dendrogram but are more distant from each other than BoNT/E and F. TeNT is an outlier, clustering with BoNT/B and BoNT/G but it is very distantly related from either of these or any other BoNT protein sequence. NTNH is the most distant protein included in the dendrogram and roots the tree.

3.1.1.2. In silico investigation of NTNH

The amino acid sequence of NTNH (A7FS63) from *C. botulinum* A1 ATCC 19397 was compared against the NCBI non-redundant protein database using BLASTp

(Table 9). The proteins with the highest similarity to NTNH/A1 were other NTNH proteins. NTNH proteins from different subtypes showed higher similarity to NTNH/A1 than the different BoNT subtypes did to BoNT/A1 NTNH/A5 was the most similar, with a similarity score of 0.98. NTNH/E and NTNH/B had similarity scores of 0.90 and 0.83 respectively. NTNH/A2, A3, A4 and F had similarity scores of 0.76, 0.76, 0.76 and 0.75 respectively. NTNH C and D had similarity scores of 0.66 and 0.65 respectively. BoNT/E was the most similar BoNT sequence compared with NTHT, with a score of 0.21. TeNT showed minimal similarity with NTNH/A1, with a similarity score of only 0.06.

A dendrogram representing the relationship between the protein sequences that showed significant similarity to NTNH/A1 was derived (Figure 16). NTNH/C and NTNH/D show strong similarity, NTNH B is most similar to C and D while NTNH/A1 and NTNH/A5 are halfway between NTNH/B and the cluster of NTNH/A2, A3, A4 and F. NTNH/E is most closely related to the NTNH sequences from the other OrfX encoding strains (A2, A3, A4 and F). BoNT/E and TeNT are both very distant from the NTNH sequences.

The NTNH sequence from each type/subtype in Table 9 was analysed using Interproscan to identify any conserved protein domains present (Figure 15). All NTNH types and subtypes had the same three functional domains in an identical arrangement along the length of the protein sequence. The three domains were – a clostridial neurotoxin zinc protease (bontoxylisin) domain (IPR000395), a clostridium neurotoxin receptor binding N-terminal domain (IPR012928) and a non-toxic non-haemagglutinin C-terminal domain (IPR013677). The bontoxylisin domain extends from the first amino acid of the N-terminus to approximately residue 446 of the 1191 amino acid NTNH/A1 protein sequence. Following the

bontoxylisin domain there is then a region with no protein domains annotated and this spans 399 amino acid residues from amino acid 446 to 845. This region is followed by the 208 amino acid neurotoxin receptor binding N-terminal that spans amino acid residues 845-1053. This region is also annotated as a lectin domain (IPR008985), involved in binding to complex carbohydrates such as gangliosides. The final domain is the Non-toxic non-haemagglutinin C-terminal domain that spans amino acid residues 1053-1196.

3.1.1.3. *In silico* investigation of HA70

The amino acid sequence of HA70 from C. botulinum A1 ATCC 19397 (A7FS58) was compared with the entire NCBI non-redundant protein database using BLASTp. The proteins with the highest similarity to HA70/A1 were other HA70 proteins. The most similar protein to HA70/A1 was HA70/A5 with a similarity score of 0.98, HA70/B also showed strong similarity to HA70/A1 with a score of 0.97. The two other subtypes associated with HA type toxin gene clusters are C and D. HA70/C had a similarity score of 0.70 while HA70/D had a score of 0.67. In addition to highly significant matches against other HA70 proteins (all E-values = 0, Table 10) BLASTp returned a match against *C. perfringens* enterotoxin with an E-value of 0.016, which is above the 0.01 cut off for significant matches. However, when the alternative sequence similarity detection tool HMMer (http://hmmer.janelia.org/) was used to analyse HA70/A1, there was a significant match (E-value = 2x10-6) between HA70/A1 and the *C. perfringens* enterotoxin. There is a low similarity score of 0.13, derived from an amino acid identity of 58% over 23% of the protein length indicating that perhaps the proteins share a functional domain.

A dendrogram representing similarity between proteins that were similar to the HA70/A1 amino acid sequence was derived (Figure 18). The HA70/A1, A5 and B proteins were all very closely related with HA70/C and HA70/D being relatively distant. The *C. perfringens* enterotoxin is very distant from the HA70 proteins.

Interproscan analysis of the HA70 types/subtypes showed that all the types/subtypes had only one domain identified – the *C. perfringens* enterotoxin domain (IPR003897). Two separate sections of the HA70 sequence were annotated with the *C. perfringens* enterotoxin domain (Figure 17), the first section spans 50-170 amino acid residues. The second domain spans from 245-455 amino acids of the 626 amino acid protein. This is in contrast to the annotation of the *C. perfringens* enterotoxin itself, which is annotated with the same domain, but in a single span covering amino acid residues 60-197 of the 316 amino acid protein (Figure 17).

3.1.1.4. *In silico* investigation of HA17

The amino acid sequence of HA17 from *C. botulinum* A1 ATCC 19397 (A7FS59) was compared against the entire NCBI non-redundant protein database using BLASTp (Table 11). The only proteins with significant similarity to HA17/A1 were other HA17 proteins. The protein with the highest similarity score compared with HA17/A1 was HA17/B with a score of 0.98, HA17/A5 was also highly similar with a score of 0.97. HA17 proteins from *C. botulinum* C and D both had similarity scores of 0.63 compared with HA17/A1.

A dendrogram representing the similarity of proteins that showed similarity to HA17/A1 was derived (Figure 20). Two clusters, one consisting of HA17/A1, A5 and B and the other consisting of HA17/C and D are evident.

The entire length of HA17/A1 shows homology with the Ricin B lectin domain (IPR000772), this domain is found in a variety of molecules including plant and bacterial AB-toxins, glycosidases and proteases (Figure 19). They are implicated in binding to sugars such as galactose and lactose.

3.1.1.5. *In silico* investigation of HA33

The amino acid sequence of HA33 (A7FS60) from *C. botulinum* A1 ATCC 19397 was compared against the entire NCBI non-redundant protein database using BLASTp. The proteins with the highest similarity to HA33/A1 were other HA33 proteins. HA33/B had a similarity score of 0.97, HA33/A5 had a similarity score of 0.91 while HA33/C and HA33/D both had similarity scores of 0.38. Two non-*C. botulinum* proteins that were similar to HA33/A1 were a mosquitocidal protein (Q03988) from *Lysinibacillus sphaericus* and Piersin-4 (C6L2F6), an apoptosis inducing protein from *Aporia crataegi* (Black-veined White butterfly).

A dendrogram was derived from the amino acid sequences of the proteins similar to HA33/A1 (Figure 22). The HA33 proteins fell into two clusters, one of HA33/B, A1 and A5 and the other of HA33/C and D. The two non-*C. botulinum* proteins that were similar to HA33, the mosquitocidal protein from *L. sphaericus* and the piersin

protein from *A. crataegi* showed very little similarity with either the HA33 proteins or each other, indicating a separate evolutionary history.

Interproscan analysis of HA33/A1 shows that it contains a Ricin B lectin domain (IPR000772). This domain is common in many plant and bacterial AB toxins, including the mosquitocidal protein and Piersin-4 identified as similar to HA33/A1 by BLASTp. It is likely that the presence of this domain explains the BLASTp hit between HA33 and the non-*C. botulinum* proteins detailed above. This domain is also present in HA17 (Figure 21 & Figure 19).

Table 8: Proteins with significant sequence similarity to BoNT/A1. E-value, identity and coverage all calculated by BLASTp. The similarity score was calculated as the product of the identity and coverage.

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity score
BoNT/A5	C7BEA8	0	98.9%	97.1%	0.96
BoNT/A2	E5F1I1	0	95.5%	90.0%	0.86
BoNT/A4	C3KS13	0	94.2%	89.4%	0.84
BoNT/A3	Q3LRX9	0	92.9%	84.9%	0.79
BoNT/F	D2KHS9	6.4x10-277	63.5%	42.9%	0.27
BoNT/E	Q54A79	7.6x10-260	65.6%	41.6%	0.27
BoNT/G	Q60393	1.1x10-256	60.6%	39.8%	0.24
BoNT/B	Q1WA38	3.1x10-254	61.7%	40.5%	0.25
TeNT	P04958	1.3x10-201	58.9%	35.4%	0.21
BoNT/D	P19321	7.9x10-201	59.9%	35.4%	0.21
BoNT/C	Q5DW55	2.2x10-185	58.7%	34.6%	0.20
NTNH	A8Y860	5.9x10-15	55.1%	31.7%	0.17

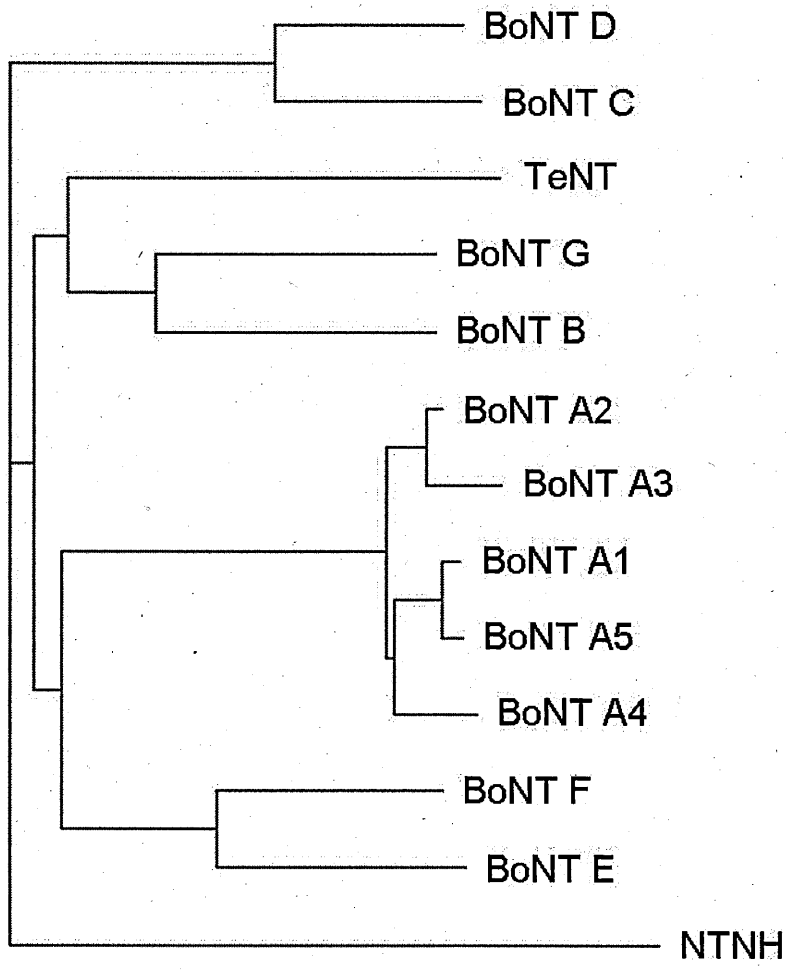
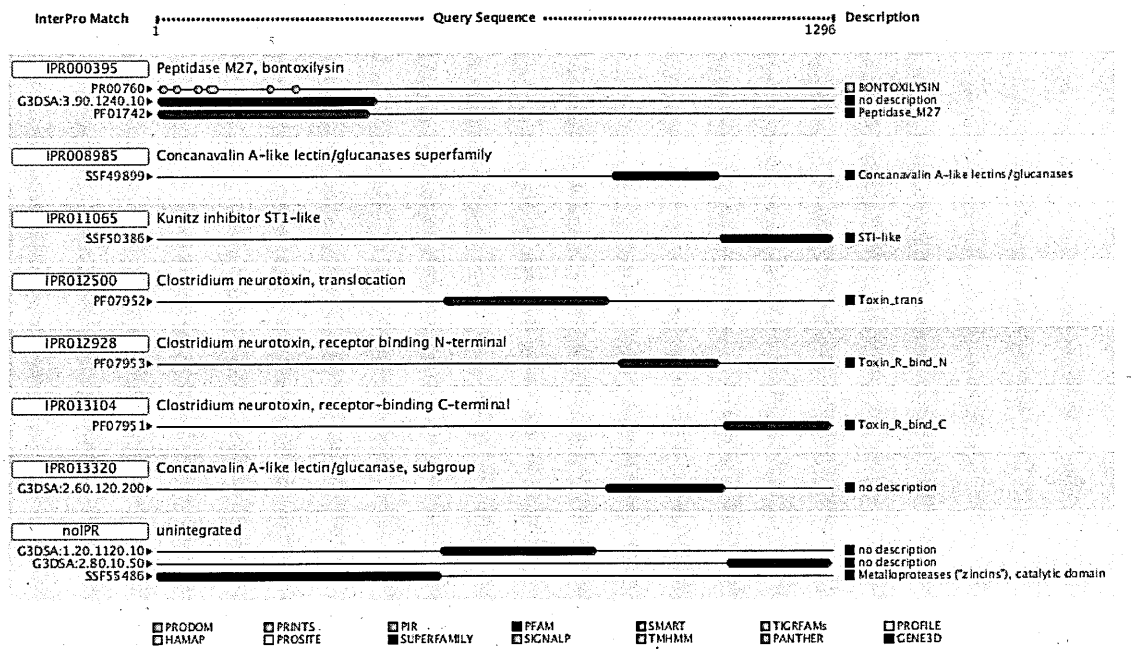


Figure 13: Dendrogram representing amino acid sequence similarity between BoNT/A-G, TeNT and NTNH. NTNH roots the dendrogram as it's sequence shows the least similarity with BoNT/A.

(A)



(B)

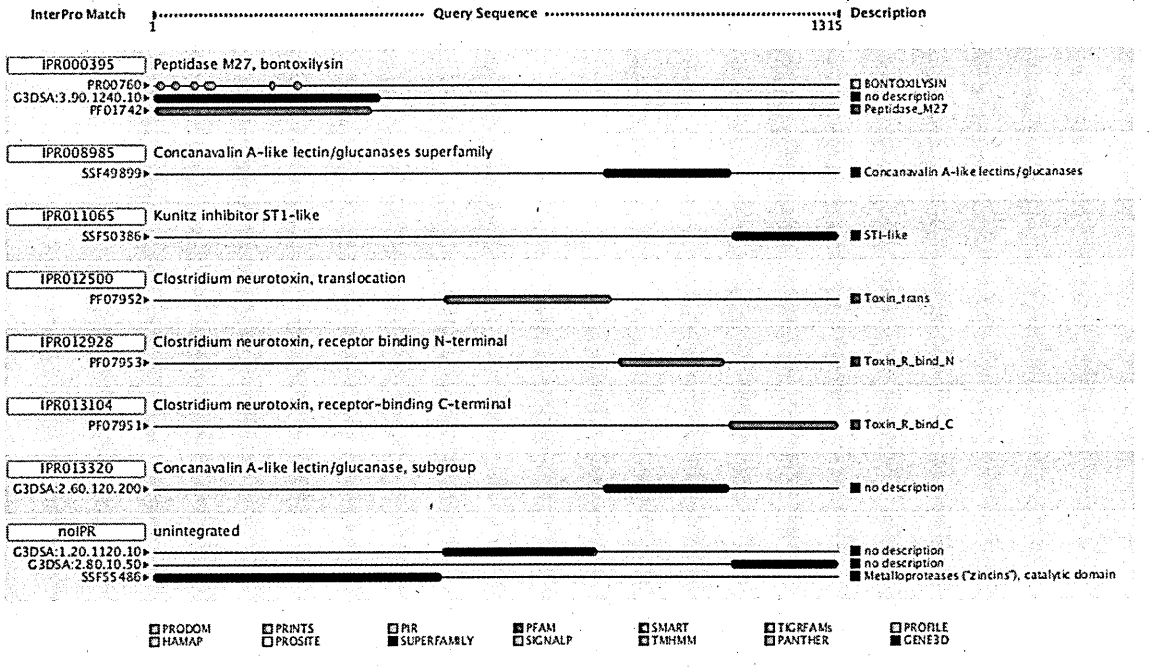


Figure 14: Protein domains identified in (A) BoNT/A1 and (B) TeNT by Interproscan. In both panels the dotted line at the top is the protein, the location and length of domains identified by Interproscan can be seen beneath the query sequence. BoNT and TeNT have identical Pfam domains in an identical arrangement.

Table 9: Proteins with significant sequence similarity to NTNH/A1. E-value, identity and coverage all calculated by BLASTp. The similarity score was calculated as the product of the identity and coverage.

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
NTNH/A5	E8ZMV9	0	98%	100%	0.98
NTNH/E	C4IHM0	0	90%	100%	0.90
NTNH/B	C3KSE1	0	83%	100%	0.83
NTNH/A3	B1L2G4	0	76%	100%	0.76
NTNH/A2	C1FUH8	0	76%	100%	0.76
NTNH/A4	C3KS14	0	76%	100%	0.76
NTNH/F	A7GBG3	0	75%	100%	0.75
NTNH/D	C5VU80	0	66%	100%	0.66
NTNH/C	B1BE43	0	66%	99%	0.65
BoNT/E	Q9K395	4x10-29	23%	92%	0.21
TeNT	TETX_CLOTE	2x10-20	28%	22%	0.06

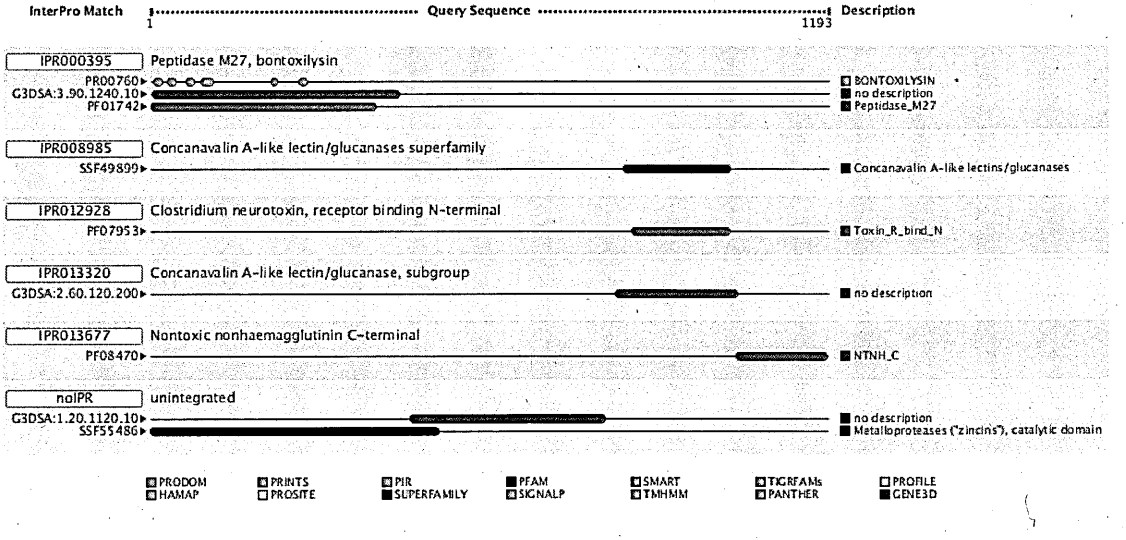


Figure 15: Protein domains identified in NTNH/A1 by Interproscan. These same protein domains were idenitified in every NTNH type/subtype analysed.

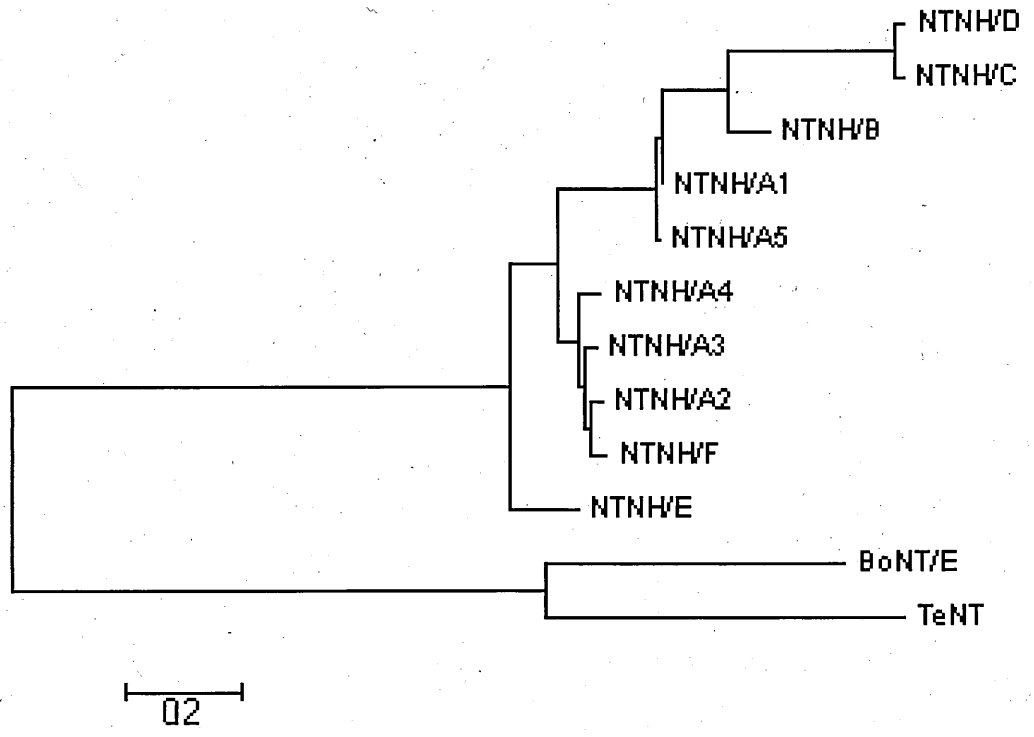
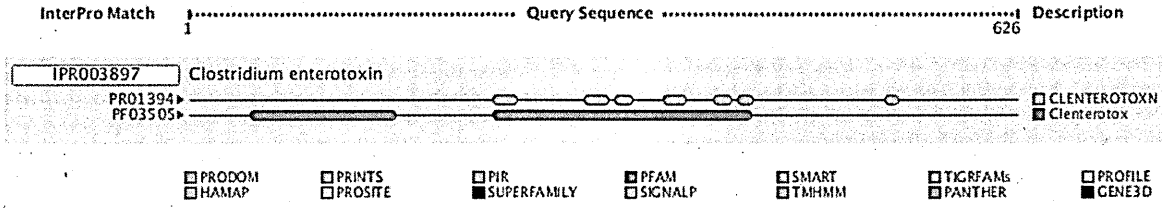


Figure 16: Dendrogram representing the relationship between proteins that showed significant similarity to NTNH/A1

Table 10: Proteins with significant sequence similarity to HA70. E-value, identity and coverage all calculated by BLASTp. The similarity score was calculated as the product of the identity and coverage.

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
HA70/A5	E8ZMV5	0	0.98	1	0.98
HA70/B	Q33CP8	0	0.97	1	0.97
HA70/C	B1BE46	0	0.7	1	0.7
HA70/D	C5VU77	0	0.68	0.98	0.67
C. perfringens enterotoxin	F7J0F5	BLASTp - 0.016	0.58	0.23	0,13
C. perfringens enterotoxin	F7J0F5	HMMer - 2x10-6	0.58	0.23	0.13

(A)



(B)

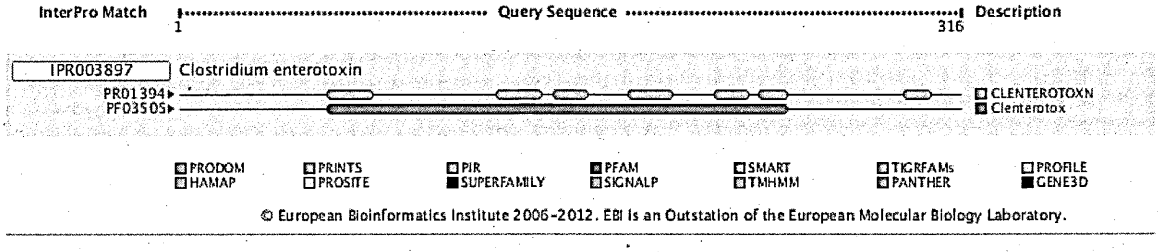


Figure 17: Protein domains identified by Interproscan in (A) HA70 (B) *C. perfringens* enterotoxin.

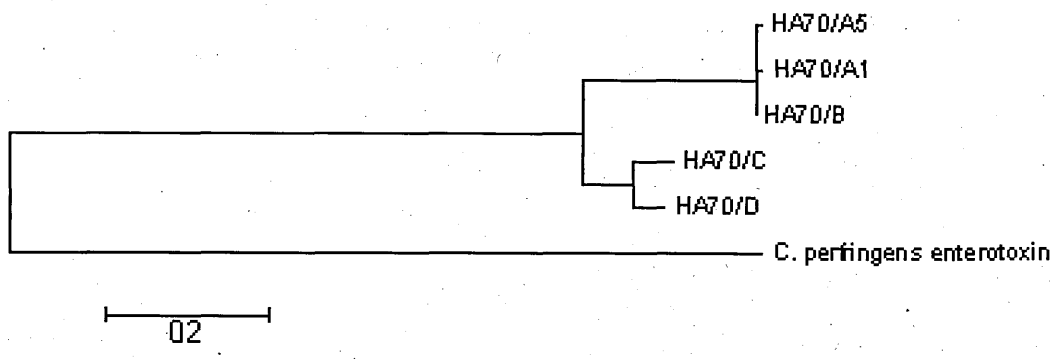


Figure 18: Dendrogram representing the relationship between proteins that showed significant similarity to HA70/A1

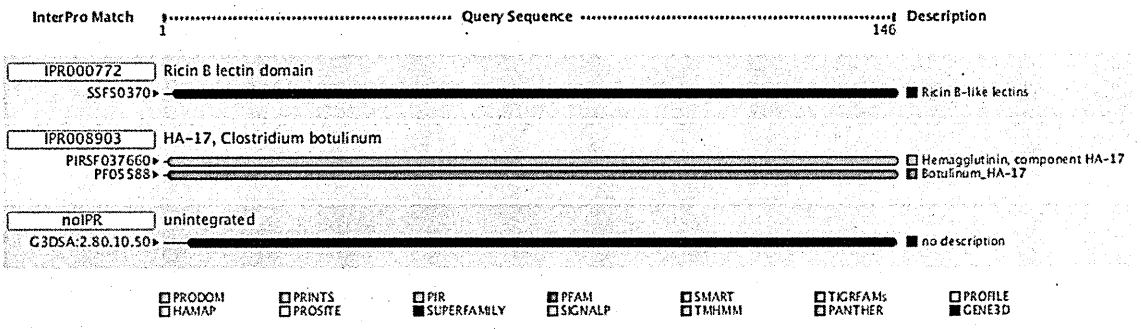


Figure 19: Protein domains identified in HA17 by Interproscan

Table 11: Signficiant matches to HA17 identified from BLASTp and HMMer analysis

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
HA17/B	Q45841	3x10-97	98%	100%	0.98
HA17/A5	E8ZMV6	6x10-96	97%	100%	0.97
HA17/C	B1BE45	3x10-52	63%	100%	0.63
HA17/D	Q9ZX78	3x10-52	63%	100%	0.63

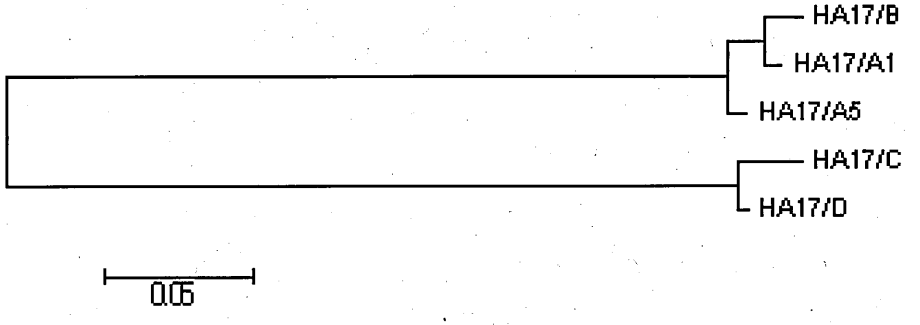


Figure 20: Dendrogram representing the relationship between proteins that showed significant similarity to HA17/A1

Table 12: Significant matches to HA33 identified by BLASTp analysis

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
HA33/B	C3KSE3	0	97%	100%	0.97
HA33/A5	E8ZMV7	0	91%	100%	0.91
HA33/C	B1BE44	2E-46	40%	94%	0.38
HA33/D	C5VU79	2E-46	40%	94%	0.38
Mosquitocidal toxin	Q03988	8E-16	33%	96%	0.32
Piersin-4	C6L2F6	0.0000002	30%	76%	0.23

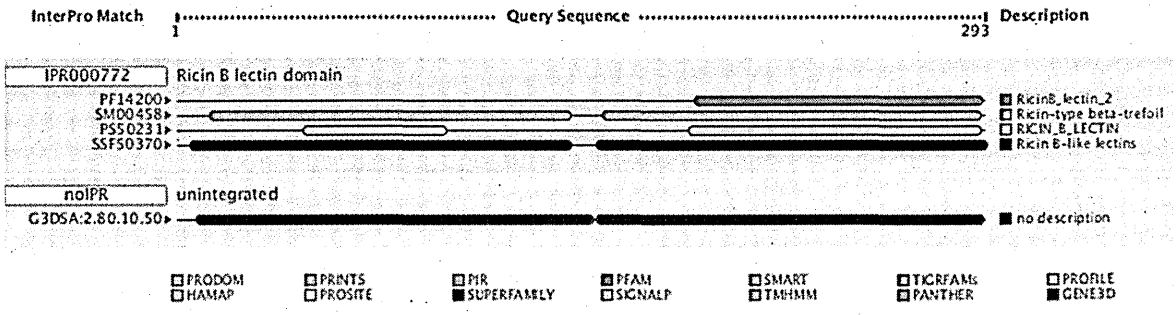


Figure 21: Protein domains identified in HA33 by Interproscan

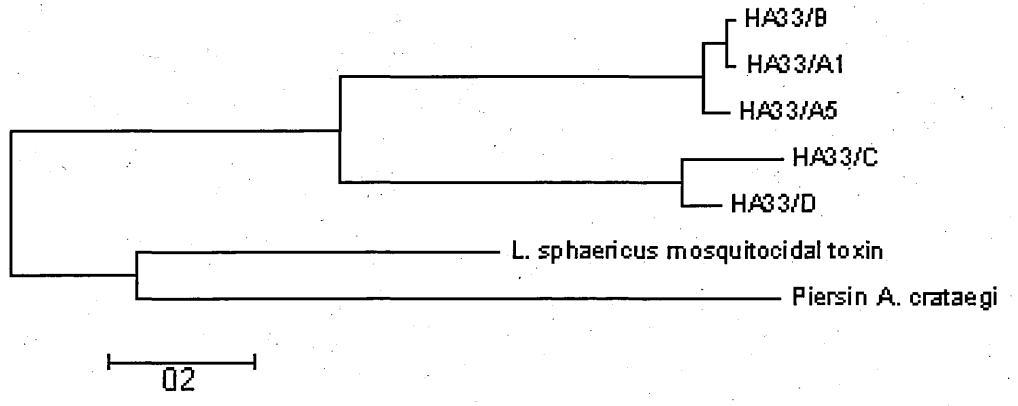


Figure 22: Dendrogram representing the relationship between proteins that showed significant similarity to HA33/A1

3.1.1.6. *In silico* investigation of ORFX1

The amino acid sequence of OrfX1 (B1L2G1) from *C. botulinum* A3 NCTC 2012 was compared against the entire NCBI non-redundant protein database using BLASTp. The proteins with the highest similarity to OrfX1/A3 were other OrfX1 proteins. The most similar protein was OrfX1/A2 with a similarity score of 0.97, OrfX1/F and OrfX1/A4 were also very similar with scores of 0.94 and 0.88 respectively. OrfX1/E showed less similarity to OrfX1/A3, with a similarity score of 0.72. There was also a *Paenibacillus dendritiformis* protein that showed a low but significant level of similarity to OrfX1/A3. This protein was annotated as an OrfX1 protein and had a similarity score of 0.25 compared with OrfX1/A3. Interproscan identified no conserved domains in OrfX1/A3. Due to the presence of a large number of non-*C. botulinum* proteins that show similarity to the OrfX proteins, dendrograms representing the similarity of these proteins will be presented and discussed in further detail in section 3.1.2.

3.1.1.7. *In silico* investigation of ORFX2

The amino acid sequence of OrfX2 from *C. botulinum* A3 NCTC 2012 (B1L2G0) was compared against the entire NCBI non-redundant protein database using BLASTp. The proteins with the highest similarity to OrfX2/A3 were other OrfX2 proteins. The most similar OrfX2 protein to OrfX3/A3 was OrfX2/A1 followed by OrfX2/A2 and OrfX2/F with similarity scores of 0.96, 0.95 and 0.95 respectively. OrfX2/A4 and OrfX2/E had similarity scores of 0.78 and 0.51 compared to OrfX2/A3.

There are multiple proteins similar to OrfX2/A3 from organisms not known to encode the botulinum neurotoxin. In many cases these proteins are annotated as OrfX2, likely because annotation of coding sequences is often based on similarity to previously identified genes. The most similar protein from a non-BoNT encoding organism is a coding sequence annotated as OrfX2 from *Paenibacillus dendritiformis*, which has a similarity score of 0.34 compared with OrfX2/A3. The same organism also encodes a protein that is similar to OrfX1. Other organisms that have coding sequences annotated as OrfX2 are *Ricketsiella grylli* (similarity score of 0.15), *Erwinia amylovora* (0.14), *Erwinia tasmaniensis* (0.12), *Erwinia pyrifoliae* (0.11) and *Arsenophonus nasoniae* (0.16). There are also significant matches with coding sequences annotated as hypothetical proteins in *Pseudomonas putida* (0.1) and *Nitrobacter winogradskyi* (0.12). OrfX2/A3 also shows similarity to another BoNT complex protein, with distant but significant similarity to OrfX3/F (similarity score of 0.12) and OrfX3/A3 (0.08).

The larger number of proteins that showed similarity to the OrfX proteins compared to the HA proteins, combined with the novelty of these observations means that an analysis of their relatedness to each other (i.e. dendrograms) will be dealt with in a separate section (3.1.2.3).

When OrfX2/A3 was analysed by Interproscan a clostridium P-47 family domain (IPR010567) was identified. This family consists of several proteins from clostridia and other species; its function is unknown. All the other proteins identified as similar to OrfX2/A3 in Table X also have a P-47 family domain.

3.1.1.8. *In silico* investigation of ORFX3

The amino acid sequence of OrfX3/A3 from *C. botulinum* A3 NCTC 2012 was compared against the entire NCBI non-redundant protein database using BLASTp. The proteins with the highest similarity to OrfX3/A3 were other OrfX3 proteins from *C. botulinum.* The most similar proteins were OrfX3/A2 and OrfX3/F which both had a similarity score of 0.97 when compared with OrfX2/A3. OrfX3/A4 and OrfX3/A1 were also very similar to OrfX3/A3 with similarity scores of 0.95 and 0.93 respectively. The most distant OrfX3 protein encoded by a BoNT encoding species was OrfX3/E with a similarity score of 0.69.

Similarly to OrfX2 there were a large number of proteins similar to OrfX3 encoded by a variety of species. Again, the species with the highest similarity protein to OrfX3/A3 was P. dendritiformis that encoded a protein annotated as OrfX3 with a similarity score of 0.45 compared with OrfX3/A3. Other species with coding sequences similar to OrfX3/A3 that were annotated as hypothetical proteins include *Halomonas* sp. TD01 (similarity score 0.30), *N. winogradskyi* (0.25), *P. putida* (0.24) and *E. tasmaniensis* (0.19). *Achromobacter xylosoxidans* encoded two proteins similar to OrfX3/A3, one annotated as OrfX3 that had a similarity score of 0.24 and another, annotated as a hypothetical protein with a score of 0.15.

OrfX3/A3 shows similarity to OrfX2/A4, with a score of 0.2 and similarity to P-47 with a score of 0.15. Some of the proteins identified as similar to OrfX3 were also identified as similar to OrfX2. Hypothetical proteins from *N. winogradskyi* (Q3SU91), *P. putida* (Q88LC9) and *E. tasmaniensis* (B2VCQ8) were all identified as being similar to both OrfX2 and OrfX3. In all three cases, the non-*C. botulinum*

proteins were more similar to OrfX3 than OrfX2 as assessed by both BLASTp E-value and similarity score.

Interproscan identified the entire length of OrfX3 as containing a clostridium P-47 domain. All the other proteins identified as similar to OrfX3/A3 were also identified as having P-47 domains.

3.1.1.9. *In silico* investigation of P-47

The amino acid sequence of P-47 from *C. botulinum* A3 NCTC 2012 (B1L2G3) was compared against the NCBI non-redundant protein database using BLASTp. The proteins with the highest similarity to P-47/A3 were other C. botulinum P-47 proteins. P-47/A1 and P-47/A2 showed the highest similarity with P-47/A3 with similarity scores of 0.99 and 0.95 respectively. P-47/F, P-47/A4 and P-47/E had similarity scores of 0.8, 0.75 and 0.73 respectively.

There were multiple proteins from non-BoNT encoding species annotated as P-47 proteins or hypothetical proteins that showed significant similarity to P-47. *P. dendritiformis* encoded the most similar protein to P-47/A3 with a similarity score of 0.36, other species encoding proteins similar to P-47/A3 included *N. winogradskyi* (0.24), *E. pyrifoliae* (0.17), *P. putida* (0.17) and *R. grylli* (0.22). P-47/A3 also showed to OrfX3/F (0.16) and OrfX2 (0.15).

Interproscan identified a P-47 domain in the P-47 protein. This protein domain family is named after this protein.

3.1.1.10. *In silico* investigation of BotR

The amino acid sequence of BotR/A1 from *C. botulinum* A1 ATCC 19397 was compared against the entire NCBI non-redundant protein database using BLASTp. The most similar proteins were BotR/B and BotR/A5 that had similarity scores of 0.98 and 0.97 compared with BotR/A1. Following these proteins, TetR, the regulator of tetanus neurotoxin showed the closest similarity with BotR/A1 with a similarity score of 0.67. BotR/F, BotR/A3 and BotR/A2 all had similarity scores of 0.61 while both BotR/C and BotR/D had similarity scores of 0.5. UviA, an RNA-polymerase sigma factor protein that regulates produciton of a bacteriocin (Dupuy et al., 2005) had a similarity score of 0.25 compared with BotR/A1.

A dendrogram representing similarity between sequences that were identified as similar to BotR/A1 was derived (Figure 26). BotR/A2, A3, A4 and F form one cluster while BotR/A1, B and A5 form a distinct cluster. TetR, the regulatory protein from *C. tetani* fell in between these two groups. BotR/C and D clustered together, distantly from these other groups.

BotR contains domains representative of various protein functions and families including transcription repressing DNA binding (IPR011991), sigma factor 70 (IPR014284) and the BotR transcription regulator (IPR017622)

Table 13: Significant matches to OrfX1/A3 identified by BLASTp analysis

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
OrfX1/A2	B0FNQ6	8E-79	97%	100%	0.97
OrfX1/F	D5VUS3	3E-75	94%	100%	0.94
OrfX1/A4	C3KS17	5E-70	88%	100%	0.88
OrfX1/E	C4IHL7	1E-58	72%	100%	0.72
OrfX1/P. dendritiformis	H3SNG9	0.00000008	26%	97%	0.25

Table 14: Significant matches to OrfX2/A3 identified by BLASTp analysis

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
OrfX2/A1	C9WWY2	0	96%	100%	0.96
OrfX2/A2	Q6RI02	0	95%	100%	0.95
OrfX2/F	D5VUS2	0	95%	100%	0.95
OrfX2/A4	C3KS18	0	78%	100%	0.78
OrfX2/E	C4IHL6	0	52%	99%	0.51
OrfX2/*P. dendritiformis*	H3SNG8	1.00E-141	35%	97%	0.34
OrfX2/ *R. grylli*	A8PMG2	1.00E-23	27%	56%	0.15
OrfX2/*E. amylovora*	E5B9Q4	5.00E-22	25%	54%	0.14
OrfX2/*E. tasmaniensis*	B2VCQ8	1.00E-21	26%	48%	0.12
OrfX2/*E. pyrifoliae*	D2T7Z3	2.00E-20	24%	46%	0.11
OrfX2/*A. nasoniae*	D2TXI8	4.00E-18	22%	74%	0.16
Hypothetical protein/*P. putida*	Q88LC9	1.00E-16	25%	39%	0.10
Hypothetical protein/ *N. winogradskyi*	Q3SU91	5.00E-16	22%	53%	0.12
OrfX3/F	A7GBF7	2.00E-08	23%	54%	0.12
OrfX3/A3	B1L2F9	9.00E-08	26%	29%	0.08

Table 15: Significant matches to OrfX3/A3 identified by BLASTp analysis

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
OrfX3/A2	Q6RI03	0	97%	100%	0.97
OrfX3/F	B1QQF2	0	97%	100%	0.97
OrfX3/A4	C3KS19	0	95%	100%	0.95
OrfX3/A1	C9WWY1	0	93%	100%	0.93
OrfX3/E	C4IHL5	0	75%	92%	0.69
OrfX3/*P. dendritiformis*	H3SNG7	2.00E-131	46%	98%	0.45
Hypothetical protein/*Halomonas* sp TD01	F7SP91	2.00E-57	31%	97%	0.30
Hypothetical protein/*N. winogradskyi*	Q3SU91	2.00E-39	26%	98%	0.25
Hypothetical protein/*P. putida*	Q88LC9	7.00E-29	24%	98%	0.24
OrfX3/*A. xylosoxidans*	F7SWZ7	2.00E-23	25%	97%	0.24
Hypothetical protein/*A. xylosoxidans*	F7SWZ6	1.00E-16	28%	55%	0.15
Hypothetical protein/*E. tasmaniensis*	B2VCQ8	7.00E-14	22%	86%	0.19
OrfX2/*A. nasoniae*	D2TXI8	2.00E-12	24%	57%	0.14
OrfX2/A4	C3KS18	4.00E-10	24%	84%	0.20
p47/*C. botulinum* A3	C3KS15	4.00E-09	26%	58%	0.15

Table 16: Significant matches to P-47/A3 identified by BLASTp analysis

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
P-47/A1	C9WWY5	0	99%	100%	0.99
P-47/A2	Q6RI05	0	95%	100%	0.95
P-47/F	A7GBG1	0	80%	100%	0.80
P-47/A4	C3KS15	0	75%	100%	0.75
P-47/E	C4IHL8	0	73%	100%	0.73
P-47/*P. dendritiformis*	H3SNG6	3.00E-88	36%	99%	0.36
Hypothetical protein/*N. winogradskyi*	Q3SU90	6.00E-21	24%	99%	0.24
P-47/*E. pyrifoliae*	D2T7Z2	3.00E-20	25%	69%	0.17
P-47/*P. putida*	Q88LC8	6.00E-17	26%	67%	0.17
Hypothetical protein/*R. grylli*	A8PMG3	1.00E-16	25%	89%	0.22
Hypothetical protein/*E. amylovora*	E5B9Q3	2.00E-16	24%	56%	0.13
Hypothetical protein/*E. tasmaniensis*	B2VCQ7	2.00E-15	26%	56%	0.15
OrfX3/F	B1QQF2	3.00E-10	26%	61%	0.16
P-47/*A. nasoniae*	D2TXI9	5.00E-10	24%	59%	0.14
P-47/*Halomonas* sp. TD01	F7SP89	1.00E-09	23%	93%	0.21
OrfX2/A3	B1L2G3	1.00E-04	27%	54%	0.15

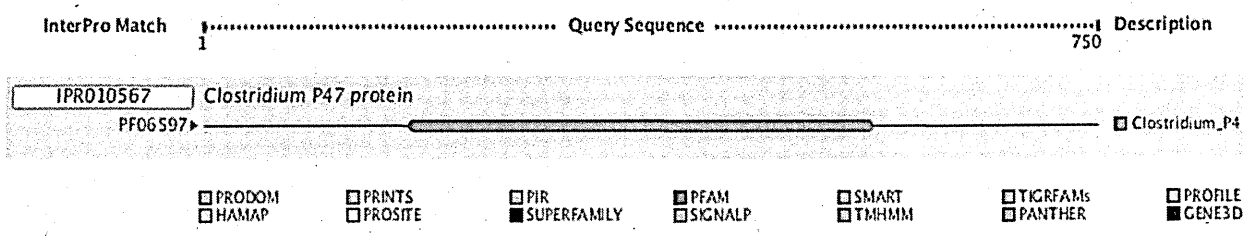


Figure 23: Protein domains identified by Interproscan in OrfX2

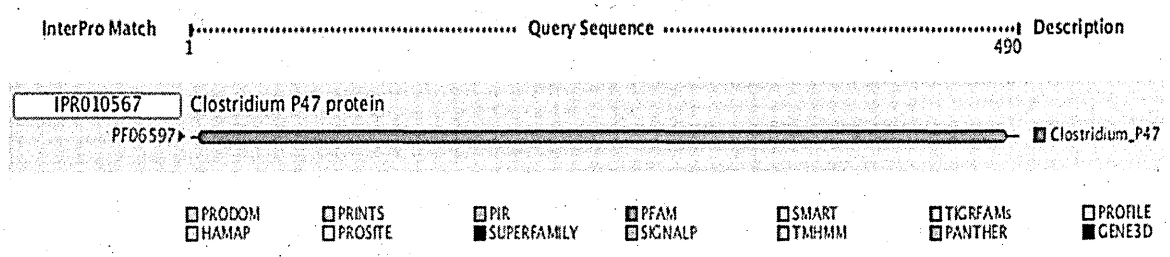


Figure 24: Domains identified in OrfX/3 by Interproscan analysis

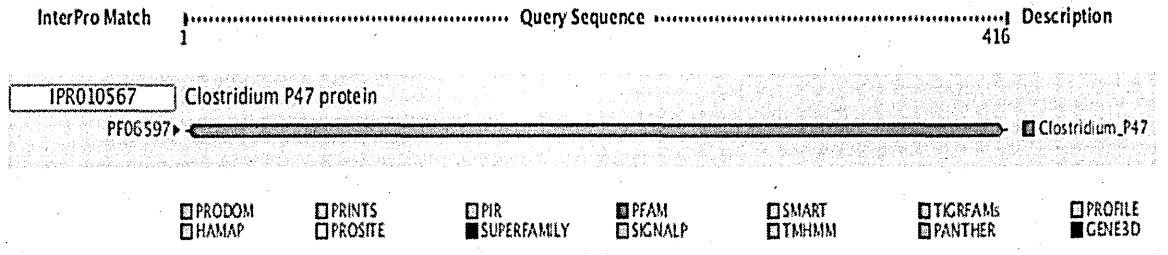


Figure 25: Domains in P-47 identified by Interproscan

Table 17: Significant matches to BotR/A1 identified by BLASTp analysis

Protein match	Uniprot ID	E-value	Identity	Coverage	Similarity
BotR/B	B1INP7	5.00E-117	98%	100%	0.98
BotR/A5	E8ZMV8	3.00E-115	97%	100%	0.97
TetR/*C. tetani*	Q899V5	9.00E-79	67%	100%	0.67
BotR/F	D5VUS4	1.00E-64	61%	100%	0.61
BotR/A3	B1L2G2	3.00E-64	62%	98%	0.61
BotR/A2	C1FUH6	6.00E-64	61%	100%	0.61
BotR/C	Q38195	1.00E-37	50%	100%	0.5
BotR/D	Q9ZWV5	2.00E-37	50%	100%	0.5

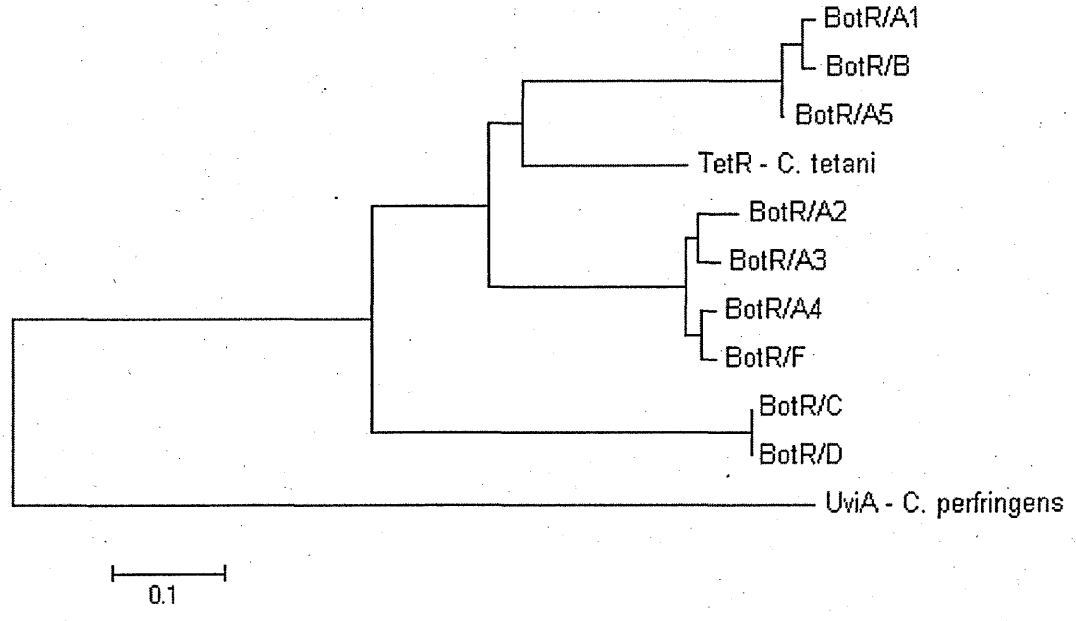


Figure 26: Dendrogram representing the relationship between proteins that showed significant similarity to BotR/A1

3.1.2. Identification of P-47 family gene clusters in non-*C. botulinum* species and their association with genes encoding putative toxin proteins

Four proteins typify the OrfX toxin complex type – OrfX1, OrfX2, OrfX3 and P-47. Interproscan identified three of these four proteins (OrfX2, OrfX3 and P-47) as containing 'P-47' domains. They also show significant homology to one another as determined by BLASTp analysis; these proteins are therefore referred to as P-47 family proteins.

Multiple P-47 family homologues were identified in nine non-*C. botulinum* species. The genomic organisation of these homologues was investigated to determine whether they are encoded in clusters, similarly to the *C. botulinum* P-47 family proteins. The nine non-*C. botulinum* species were *Ricketsiella grylli, Paenibacillus larvae, Erwinia tasmaniensis, Arsenophonus nasoniae, Halomonas* sp TD01, *Paenibacillus dendritiformis, Nitrobacter winogradskyi*, *Achromobacter xylosoxidans* and *Pseudomonas putida.* In all nine species, the P-47 family proteins were clustered together in a broadly similar fashion to the *C. botulinum* P-47 family proteins, although there was variation in how many homologues were present and their precise organisation. Furthermore, in 6 of the 10 strains examined (including *C. botulinum*), genes encoding known or putative toxins were co-localised with P-47 family proteins.

3.1.2.1. Analysis of non-*C. botulinum* species in which P-47 family clusters are associated with putative toxin proteins

Paenibacillus larvae

P. larvae is an anaerobic, endospore forming bacteria found in a range of environments such as soil, water and insect larvae, it causes American foulbrood in honeybees. P. larvae encodes a cluster of three P-47 family encoding genes. Two of these are degraded by frame shift and stop codon insertions (Figure 27 & Table 18). As these proteins are degraded they are not present in the NCBI or Uniprot databases, this explains why there were no matches to these proteins when these databases were searched for proteins similar to the OrfX proteins. Two of the P. larvae CDSs are annotated as OrfX2 and OrfX3 due to a strong similarity with the C. botulinum OrfX2 and OrfX3. The P. larvae OrfX2 and OrfX3 encoding genes also show synteny with the C. botulinum genes.

The P-47 cluster also encodes a degraded protein (annotated as two open reading frames - PlarlB_020100022873/degraded feature 3) that shows similarity to the clostridial C2 toxin and Bacillus anthracis Protective Antigen (PA) as well as close similarity with a protein annotated as Protective Antigen in P. dendritiformis. These four proteins (P. larvae, P. dendritiformis and B. anthracis PA and the clostridial C2 toxin) all have PA14 (IPR011658) and bacterial exotoxin B (IPR003896) domains. The PA14 domain is thought to play have a role in binding of the Protective Antigen and is present in many bacterial enzymes and toxins. The bacterial exotoxin B domain is characteristic of the binding subunit of an AB toxin.

Ricketsiella grylli

Ricketsiella grylli is an intracellular pathogen of insects, including crickets. R. grylli encodes two P-47 family genes (Figure 28 & Table 19). One gene encodes a protein similar to OrfX2 (similarity score 0.24) and the other a protein similar to P-47 (0.23). Two other proteins of interest encoded in the *R. grylli* P-47 family gene cluster are a 17-kDa protein also identified in P-47 family gene clusters in A. nasoniae and *E. tasmaniensis* which has been named P-17. This protein doesn't show similarity to any well characterised proteins and no known domains were identified in it's sequence on Interproscan analysis. The *R. grylli* P-47 family cluster also encodes a Shiga-toxin A-chain homologue immediately upstream of the cluster, on the opposite strand, an arrangement similar to the organisation of the P-47 family gene cluster and the BoNT gene in *C. botulinum*.

Erwinia tasmaniensis

E. tasmaniensis is a non-pathogenic species associated with plants. It encodes two P-47 family genes (Figure 29 & Table 20), one protein that is most similar by BLASTp E-value to OrfX2 and another that is most similar to P-47. However, both these proteins also show approximately equivalent similarity to other *C. botulinum* P-47 family proteins. A CDS similar to the P-17 protein identified in *R. grylli* and *A. nasoniae* was also identified. *E. tasmaniensis* encodes a toxin within the P-47 family gene cluster; a nematicidal protein homologue identified in bacteriophage and fungi. There was also a hypothetical protein in the P-47 family cluster that was identified in the same bacteriophage as the nematicidal protein.

Arsenophonus nasoniae

A. nasoniae encodes a cluster of three P-47 family CDSs (Figure 30 & Table 21). Two of these CDSs show highest similarity with OrfX2 (similarity scores of 0.31

and 0.17) and one has closest similarity to P-47 (0.15). However, these proteins also show a comparable level of similarity to other P-47 family protein sequences. The *A. nasoniae* P-47 cluster also encodes a protein similar to the P-17 hypothetical protein in *R. grylli* and *E. tasmaniensis*. Also encoded within the A. nasoniae P-47 cluster are two putative nematicidal proteins which show similarity to the nematicidal protein identified in a P-47 cluster in *E. tasmaniensis*. These two proteins show similarity to putative toxin proteins in *Yersinia enterocolitica* among others.

***Halomonas* sp. TD01**

Halomonas sp. TD01 encodes a cluster containing three P-47 family CDSs (Figure 31 & Table 22). Two of the genes encode proteins that show the highest similarity with OrfX3 with similarity scores of 0.31 and 0.11 respectively, the third CDS encodes a protein that is most similar to P-47 (0.2). However, these proteins also show a comparable level of similarity to other P-47 family protein sequences. *Halomonas* sp. TD01 also encodes two nematicidal proteins alongside the P-47 family cluster. When these proteins are compared to each other they have similarity score of 0.49/0.48. Both of these proteins also show similarity to AidA nematicidal proteins from *Ralstonia salanaceaerum* and *Burkholderia cepacia* (plant pathogens) as well a hypothetical protein from *Xenorhabdus nematophila* (insect pathogen).

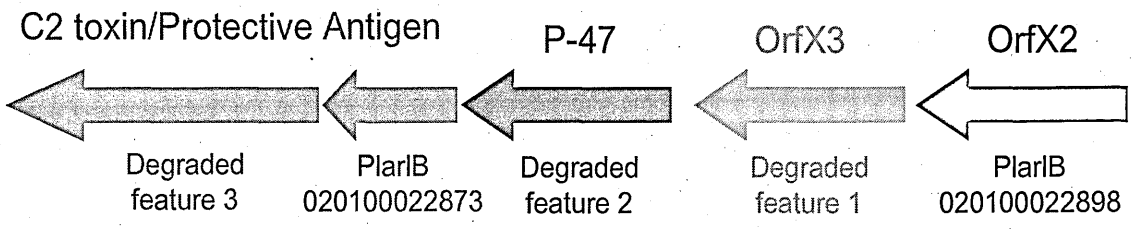


Figure 27: The *P. larvae* P-47 family gene cluster. There are three P-47 family CDSs, two of which are degraded by stop codon insertion and frame shift mutations. There is also a partially degraded C2 toxin/Protective Antigen homologue.

Table 18: BLASTp matches to proteins from the *P. larvae* P-47 family cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
PlarlB_020100022898	OrfX2	1.00E-145	36%	99%	0.36	YP_001715698.1
	OrfX3	3.00E-15	25%	37%	0.09	YP_001715697.1
Degraded feature 1	OrfX3	6.00E-143	46%	97%	0.45	YP_001715697.1
	OrfX2	4.00E-14	22%	86%	0.19	YP_001715698.1
Degraded feature 2	P-47	5.00E-92	36%	91%	0.33	YP_001715701.1
	OrfX2	1.00E-08	24%	52%	0.12	YP_001715698.1
	OrfX3	9.00E-08	22%	83%	0.18	YP_001715697.1
PlarlB_020100022873/degraded feature 3	Protective antigen (*P. dendritiformis*)	0.00E+00	74%	98%	0.73	ZP_09677307
	Iota toxin component Ib (*C. perfringens*)	1.7x10-79	30%	82%	0.25	YP_004670323
	C2 toxin, component II (*C. botulinum* D)	2.1x10-71	31%	62%	0.19	YP_003034266
	Iota toxin component Ib (*B. thurigenesis*)	2.7x10-62	28%	80%	0.22	EEM56627
	Protective antigen (*B. anthracis*)	7.9x10-49	29%	67%	0.19	YP_002811596

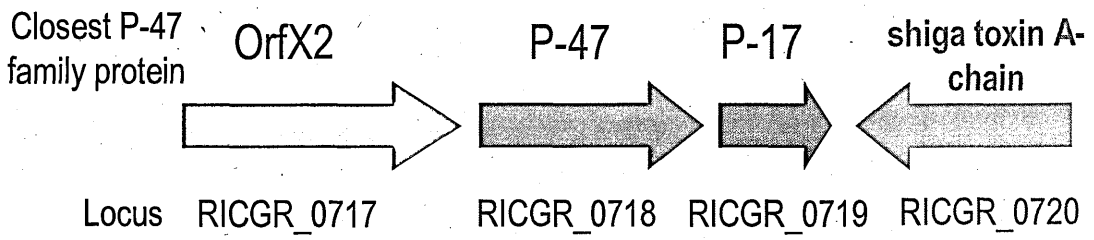


Figure 28: The *R. grylli* P-47 family gene cluster. There are two P-47 family CDSs, one encoding a novel protein conserved in other species (P-17), and a Shiga-toxin A-chain homologue.

Table 19: BLASTp matches to proteins from the *R. grylli* P-47 family cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
RICGR_0717	OrfX2	1.00E-29	25%	97%	0.24	YP_001715698.1
	OrfX3	9.00E-09	23%	51%	0.12	YP_001715697.1
RICGR_0718	P-47	2.00E-20	25%	90%	0.23	YP_001715701.1
	OrfX3	2.00E-08	24%	62%	0.15	YP_001715697.1
	OrfX2	0.024	23%	44%	0.1	YP_001715698.1
RICGR_0719	Hypothetical protein (*A. nasoniae*)	1.30E-30	48%	93%	0.45	CBA72108.1
	Hypothetical protein (*E. tasmaniensis*)	8.00E-12	36%	68%	0.24	CAO98071.1
RICGR_0720	Shiga toxin 2f A subunit (*E. coli*)	4.50E-60	41%	98%	0.4	BAE79483.1

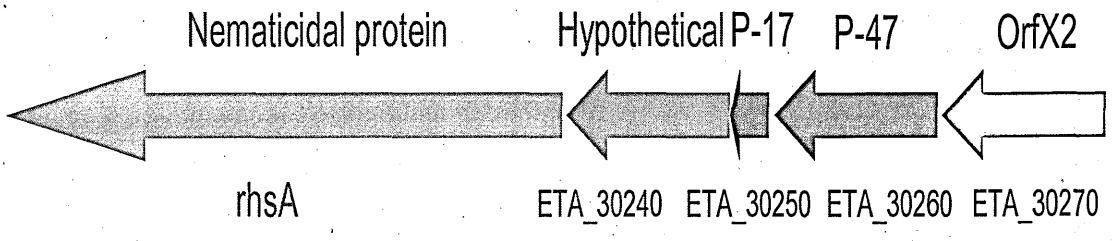


Figure 29: The *E. tasmaniensis* P-47 family cluster. There are two P-47 family CDSs, a P-17 sequence, a hypothetical CDS and a nematicidal protein homologue.

Table 20: BLASTp matches to proteins from the *E. tasmaniensis* P-47 family cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
ETA_30270	OrfX2	2.00E-25	26%	65%	0.17	YP_001715698.1
	OrfX3	5.00E-18	22%	83%	0.18	YP_001715697.1
	P-47	3.00E-05	23%	42%	0.1	YP_001715701.1
ETA_30260	P-47	3.00E-19	26%	57%	0.15	YP_001715701.1
	OrfX3	6.00E-10	23%	68%	0.16	YP_001715697.1
ETA_30250	Hypothetical proteins (*R. grylli*)	1.10E-11	36%	67%	0.24	ACJ10120.1
	Hypothetical protein (*A. nasonaie*)	9.60E-06	30%	74%	0.22	XP_002378664.1
ETA_30240	Hypothetical protein (Bacteriophage APSE-3)	5x10-67	38%	95%	0.36	ACJ10121.1
	Hypothetical protein (*Aspergillus flavus*)	6.70E-10	22%	94%	0.21	EEQ17818.1
rhsA	Putative YD-repeat toxin (Bacteriophage APSE-3)	0.00E+00	51%	87%	0.44	ZP_12869720.1
	Rhs family protein (*Yersinia intermdeia*)	0.00E+00	40%	87%	0.35	CAC19493.1
	Rhs family protein (*Yersinia enterocolitica*)	0.00E+00	40%	87%	0.35	GAA91443.1
	Nematicidal protein (*Xenoharbdus bovienii*)	1.9x10-252	36%	88%	0.32	YP_003712026.1
	RHS repeat protein (*Aspergillus kawachii*)	4.6x10-188	33%	84%	0.28	EDP45744.1
	Putative nematicidal protein (*Xenorhabdus nematophila*)	9.8x10-180	33%	82%	0.27	CBA72108.1

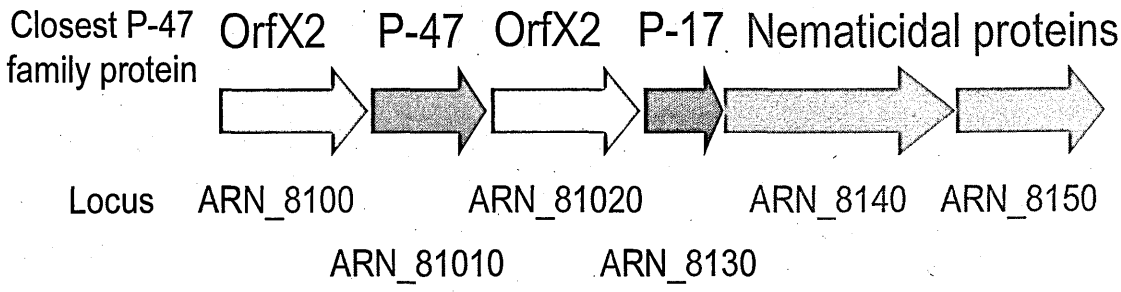


Figure 30: The *A. nasoniae* P-47 family cluster. There are three P-47 family CDSs, a P-17 homologue and two CDSs encoding nematicidal protein homologues.

Table 21: BLASTp matches to proteins from the *A. nasoniae* P-47 family cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
ARN_08100	OrfX2	1.00E-25	32%	96%	0.31	YP_001715698.1
	OrfX3	8.00E-19	25%	51%	0.13	YP_001715697.1
ARN_081010	P-47	8.00E-14	24%	63%	0.15	YP_001715701.1
	OrfX3	3.00E-10	22%	73%	0.16	YP_001715697.1
	OrfX2	0.05	45%	6%	0.03	YP_001715698.1
ARN_081020	OrfX2	4.70E-23	23%	74%	0.17	YP_001715698.1
	OrfX3	6.70E-05	23%	51%	0.12	YP_001715697.1
	P-47	0.0057	23%	46%	0.11	YP_001715701.1
ARN_08130	Hypothetical protein (*R. grylli*)	4.50E-31	48%	93%	0.45	EDP45744.1
	Hypothetical protein (*E. tasmaniensis*)	2.70E-06	30%	76%	0.23	CAO98071.1
ARN_08140	Hypothetical protein (*Photorhabdus luminsescens*)	5.40E-173	37%	99%	0.37	NP_929681.1
	Nematicidal protein 2 (*E. tasmaniensis*)	4.80E-121	30%	98%	0.29	CAO95593.1
	Rhs family protein (*Yersinia enterocolitica*)	1.80E-113	29%	99%	0.29	ZP_12869720.1
	Nematicidal protein (*Xenorhabdus bovieni*)	1.20E-104	30%	99%	0.3	CAC19493.1
	Nematicidal protein (*Xenorhabdus nematophilia*)	3.40E-99	29%	96%	0.28	YP_003712026.1
	Putative YD-repeat toxin (Bacteriophage APSE-3)	1.10E-95	29%	99%	0.29	ACJ10121.1
	RHS repeat protein (*Aspergillus kawachii*)	2.70E-90	29%	99%	0.29	GAA91443.1
ARN_08150	Hypothetical protein (*Photorhabdus luminsescens*)	4.10E-98	37%	96%	0.36	NP_929681.1
	Putative YD-repeat toxin (Bacteriophage APSE-3)	3.00E-89	29%	99%	0.29	ACJ10121.1
	Rhs family protein (*Yersinia enterocolitica*)	6.60E-81	29%	99%	0.29	ZP_12869720.1
	Nematicidal protein 2 (E. tasmaniensis)	4.00E-76	27%	98%	0.26	CAO95593.1

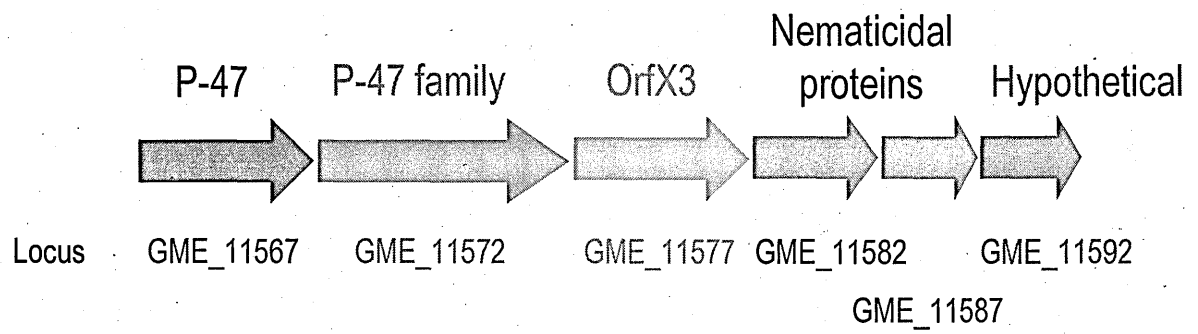


Figure 31: The *Halomonas* P-47 family cluster. There are three P-47 family CDSs and two nematicidal protein homologue.

Table 22: BLASTp matches to proteins from the *Halomonas* sp. TD01 P-47 family gene cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
GME_11567	P-47	2.00E-13	23%	88%	0.2	YP_001715701.1
	OrfX3	0.003	22%	54%	0.12	YP_001715697.1
	OrfX2	0.076	23%	40%	0.09	YP_001715698.1
GME_11572	OrfX3	3.00E-15	30%	35%	0.11	YP_001715697.1
	OrfX2	2.00E-07	28%	13%	0.04	YP_001715698.1
	P-47	6.00E-06	22%	25%	0.06	YP_001715701.1
GME_11577	OrfX3	3.00E-67	32%	96%	0.31	YP_001715697.1
	P-47	2.00E-10	21%	78%	0.16	YP_001715701.1
	OrfX2	5.00E-09	20%	52%	0.1	YP_001715698.1
GME_11582	Hypothetical protein (*Halomonas* GME_11587)	3.20E-50	50%	97%	0.49	ZP_08637335.1
	Nematicidial protein AidA (*Ralstonia solanacearum*)	8.00E-47	50%	97%	0.49	YP_003744150.1
	Nematicidial protein AidA (*Burkholderia cepacia*)	3.50E-35	38%	97%	0.37	YP_002153683.1
	Hypothetical protein (*Xenorhabdus nematophila*)	2.60E-19	31%	96%	0.3	YP_003712766.1
GME_11587	Hypothetical protein (*Halomonas* GME_11582)	3.10E-50	50%	96%	0.48	EGP19391.1
	Nematicidial protein AidA (*Ralstonia solanacearum*)	8.00E-47	50%	97%	0.49	YP_003744150.1
	Nematicidial protein AidA (*Burkholderia cepacia*)	3.50E-35	38%	97%	0.37	YP_002153683.1
	Hypothetical protein (*Xenorhabdus nematophila*)	2.60E-19	31%	96%	0.3	YP_003712766.1

3.1.2.2. Analysis of species with P-47 family proteins not associated with toxin proteins

There were four species in which the P-47 family encoding sequences were not associated with a putative toxin encoding gene - *Paenibacillus dendritiformis, Nitrobacter winogradskyi, Achromobacter xylosoxidans* and *Pseudomonas putida.*

The P-47 family gene cluster of *P. dendritiformis* encodes four P-47 family genes (
Figure 32 & Table 23). It also shows the greatest synteny to the *C. botulinum* P-47 family gene cluster – it is the only non-BoNT producing species that encodes a protein similar to OrfX1. These coding sequences were the only sequences on a contig that resulted from next-generation sequencing data of this organism, therefore, the exact genomic surroundings of this cluster are not known.

N. winogradskyi encodes four P-47 family proteins, three proteins similar to OrfX3 and one protein similar to P-47 (Figure 33 & Table 24). *A. xylosoxidans* encodes three P-47 family proteins, one P-47 family protein which shows equal similarity to both OrfX2 and OrfX3, one OrfX3 homologue and one protein similar to P-47 (Figure 34 & Table 25). *P. putida* encodes two P-47 family proteins, one protein similar to P-47 and one protein similar to OrfX3 (Figure 35 & Table 26).

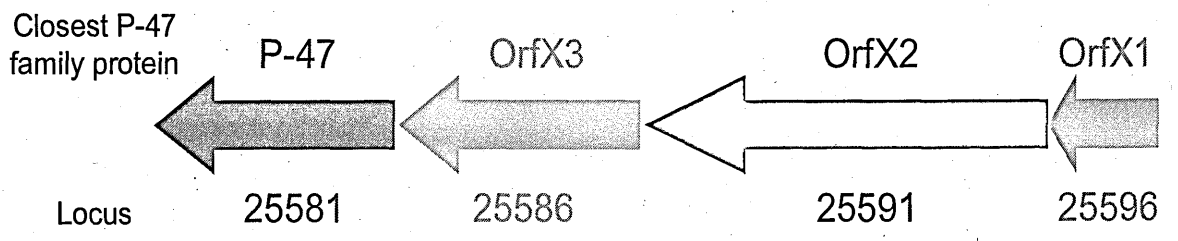


Figure 32: The *P. dendritiformis* P-47 family cluster encodes four P-47 family coding sequences.

Table 23: BLASTp matches to proteins from the *P. dendritiformis* P-47 family cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
PDENDC454_25596	OrfX1	2.30E-09	25%	97%	0.24	ACA57457.1
PDENDC454_25591	OrfX2	2.00E-138	36%	99%	0.36	ACA57564.1
	OrfX3	9.50E-12	25%	37%	0.09	ACA57304.1
PDENDC454_25586	OrfX3	2.60E-130	46%	97%	0.45	ACA57304.1
	OrfX2	1.70E-16	23%	82%	0.19	ACA57564.1
PDENDC454_25581	P-47	3.20E-81	36%	91%	0.33	YP_001715701.1

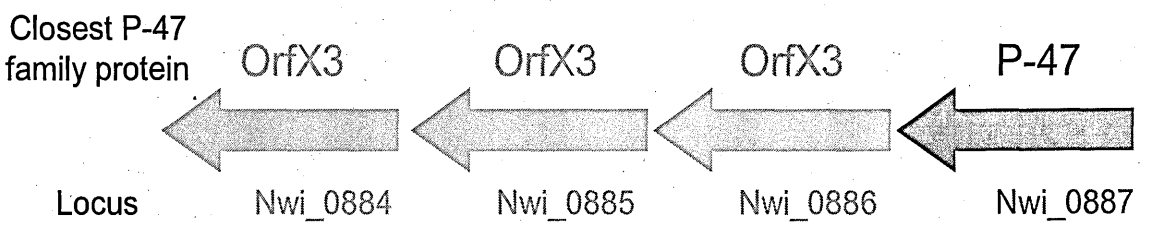


Figure 33: The *N. winogradskyi* P-47 family cluster encodes four CDSs.

Table 24: BLASTp matches to proteins from the *N. winogradskyi* P-47 family cluster.

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
Nwi_0887	P-47	1.00E-26	24%	98%	0.24	YP_001715701.1
	OrfX3	7.00E-18	22%	73%	0.16	YP_001715697.1
Nwi_0886	OrfX3	7.00E-48	27%	87%	0.23	YP_001715697.1
	OrfX2	5.00E-20	36%	87%	0.31	YP_001715698.1
	P-47	4.00E-14	21%	56%	0.12	YP_001715701.1
Nwi_0885	OrfX3	1.00E-34	23%	88%	0.2	YP_001715697.1
	OrfX2	1.00E-12	27%	37%	0.1	YP_001715698.1
	P-47	7.00E-07	22%	58%	0.13	YP_001715701.1
Nwi_0884	OrfX3	7.00E-09	20%	87%	0.17	YP_001715697.1
	P-47	0.041	20%	24%	0.05	YP_001715701.1

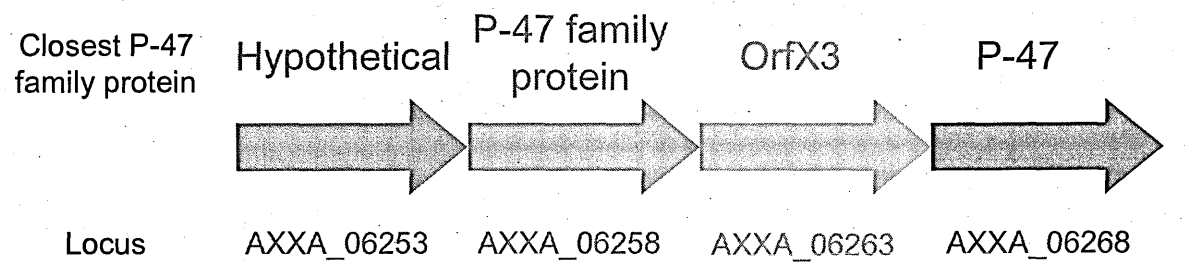


Figure 34: The *A. xylosoxidans* P-47 family cluster encodes three P-47 family coding sequences.

Table 25: BLASTp matches to proteins from the *A. xylosoxidans* P-47 family cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
AXXA_06258	OrfX3	2.00E-20	28%	49%	0.14	YP_001715697.1
	OrfX2	5.00E-17	23%	99%	0.23	YP_001715698.1
	P-47	6.00E-12	21%	47%	0.1	YP_001715701.1
AXXA_06263	OrfX3	2.00E-27	25%	98%	0.25	YP_001715697.1
	OrfX2	4.00E-12	23%	51%	0.12	YP_001715698.1
	P-47	3.00E-10	20%	51%	0.1	YP_001715701.1
AXXA_06268	OrfX3	1.00E-04	24%	22%	0.05	YP_001715697.1

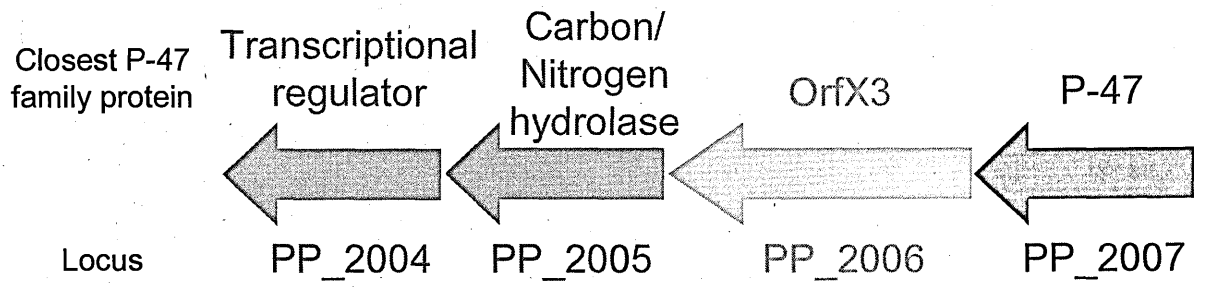


Figure 35: The *P. putida* P-47 family cluster encodes three P-47 family CDSs.

Table 26: BLASTp matches to proteins from the *P. putida* P-47 family cluster

Query locus	Match protein	Match score	Identity	Coverage	Similarity	Match id
PP_2007	P-47	2.00E-25	25%	68%	0.17	YP_001715701.1
	OrfX3	3.00E-10	21%	74%	0.16	YP_001715697.1
	OrfX2	3.00E-05	22%	49%	0.11	YP_001715698.1
PP_2006	OrfX3	5.00E-37	26%	78%	0.2	YP_001715697.1
	OrfX2	8.00E-21	22%	69%	0.15	YP_001715698.1
	P-47	1.00E-09	22%	50%	0.11	YP_001715701.1
PP_2005	Hypothetical protein (*Pseudomonas syringe*)	1.20E-137	58%	100%	0.58	YP_234760.1
PP_2004	AraC family transcriptional *regulator (Pseudomonas syringe*)	8.90E-166	74%	97%	0.72	EGH03104.1

3.1.2.3. Phylogenetic analysis of P-47 family proteins and the species encoding them

P-47 family coding sequences have been identified in nine non-*C. botulinum* species. In five of these species the P-47 family cluster is encoded alongside a putative toxin gene. Comparison of the P-47-family protein sequences from the different species will provide information on the similarity of the coding sequences that make up the clusters. Contrasting this data with phylogenetic analysis based on 16S rDNA will provide insight into the evolutionary history of the P-47-family genes.

The P-47 proteins from different species were grouped by the *C. botulinum* P-47 family protein they were most similar to (i.e. OrfX2, OrfX3 or P-47). In some cases there was not a significant difference between the highest similarity match and the second highest e.g. the *A. nasoniae* CDS Arn_08100 was grouped with the OrfX2 proteins despite having a BLASTp E-value of $1x10^{-25}$ compared with OrfX2 and an E-value of $8x10^{-19}$ compared with OrfX3. However, including many highly divergent protein sequences disrupts the process of sequence alignment so the sequences were compartmentalised. The amino acid sequences of each protein group (i.e. the OrfX2 proteins etc) were aligned using ClustalW and a pairwise similarity matrix was calculated (see appendix for full matrices). Trees were derived from the resulting matrices using Fitch-Margoliash clustering implemented by PHYLIP (http://evolution.genetics.washington.edu/phylip.html) and visualised using Dendroscope. (http://ab.inf.uni-tuebingen.de/software/dendroscope/). All trees were rooted on *Halomonas* TD01 locus 11572 as this sequence was highly divergent from the other sequences.

There were 6 non-*C. botulinum* proteins that were closest in similarity to OrfX2 and a tree representing protein distance was derived for these proteins (Figure 38). The *C. botulinum* sequences cluster together, with the BoNT/A3 and A4 associated sequences clustering more closely compared with the BoNT/E associated sequence. *P. dendritiformis* and *P. larvae* don't cluster together but both fall near to the *C. botulinum* sequences. The sequences from *E. tasmaniensis*, *A. nasoniae* and *R. grylli* all cluster together.

On the OrfX3 tree (Figure 37) a similar pattern emerges. The *C. botulinum* sequences cluster together, with the *P. larvae* and *P. dendritiformis* sequences clustering together, close to the *C. botulinum* sequences. The *P. putida, N. winogradskyi* and *A. xylsoxidans* sequences show significant difference from any other species. The three *N. winogradskyi* OrfX3 sequences which cluster together show significant sequence divergence from each other.

On the P-47 tree (Figure 38) the *C. botulinum* sequences also cluster together, with BoNT/A3 and A4 sequences showing more homology to each other than to BoNT/E. *P. larvae* and *P. dendritiformis* cluster together, close to the *C. botulinum* sequences. The sequences from *E. tasmaniensis*, *A. nasoniae* and *R. grylli* all cluster together, similarly to the OrfX2 tree. *P. putida* and *N. winogradskyi* cluster loosely together, showing significant sequence difference from any other P-47 protein.

A phylogenetic tree was derived using 16S rDNA sequences from the species which encode P-47 family proteins (Figure 39). The similarity between this tree and the P-47 family protein trees, in particular the P-47 protein tree, is notable. There are two broad clusters that correspond to the two main phyla with P-47

family genes – the Firmicutes and the Gammaproteobacteria. The group I (A3 and Ba4) and group II (E) *C. botulinum* strains cluster together while the *Paenibacillus* species are on a neighbouring branch - these species are all Firmicutes. The other broad cluster consists of *R. grylli*, *P. putida, Halomonas* TD01, *E. tasmaniensis* and *A. nasoniae* - these species are all Gammaproteobacteria. *N. winogradskyi,* an Alphaproteobacteria, and *A. xylosoxidans*, a Bacteroidetes, cluster separately from each other and the other species.

One major difference between the 16S tree and the OrfX2/P-47 protein trees is that, in the 16S tree *Halomonas* TD01 and *P. putida* cluster more closely to *E. tasmaniensis* and *A. nasoniae* than *R. grylli*. In the P-47 and OrfX2 protein trees *E. tasmaniensis, A. nasoniae* and *R. grylli* cluster together and *P. putida* and *Halomonas* TD01 show lower similarity. This suggests that the *R. grylli* P-47 family protein sequence was acquired more recently than the *P. putida/Halomonas* TD01 sequence, likely by horizontal gene transfer. There is also a large degree of synteny between the P-47 family clusters of *E. tasmaniensis*, *R. grylli* and *A. nasoniae*.

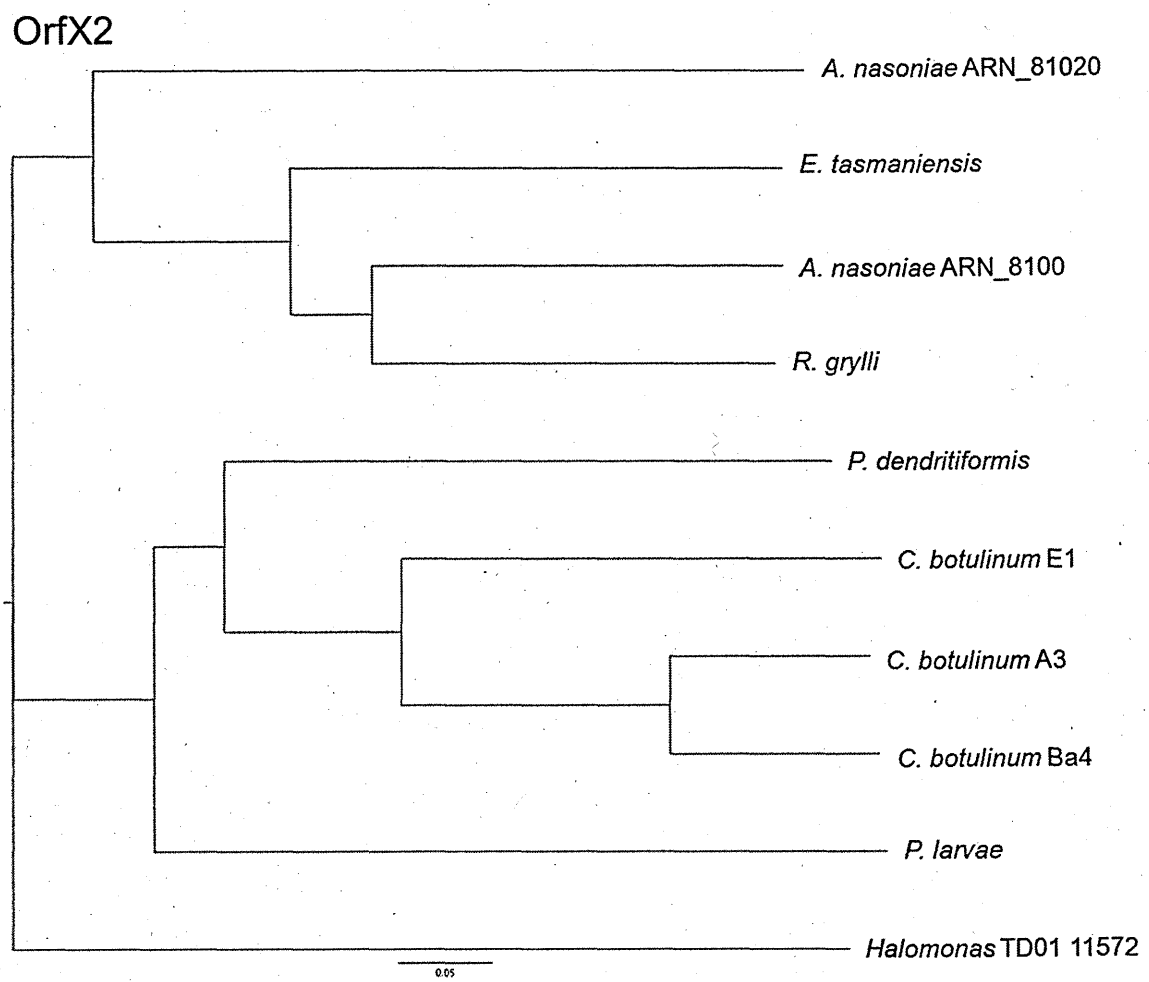


Figure 36: Neighbour joining protein distance tree of OrfX2 sequences from P-47 family encoding species. There are two main clusters – the *C. botulinum* and *Paenibacillus* species sequences cluster together, and the *A. nasoniae, E. tasmaniensis* and *R. grylli* sequences cluster together separately. *Halomonas* TD01 GME_11572 is a P-47 family protein which shows equal similarity to OrfX2 and OrfX3 and was used to root the tree.

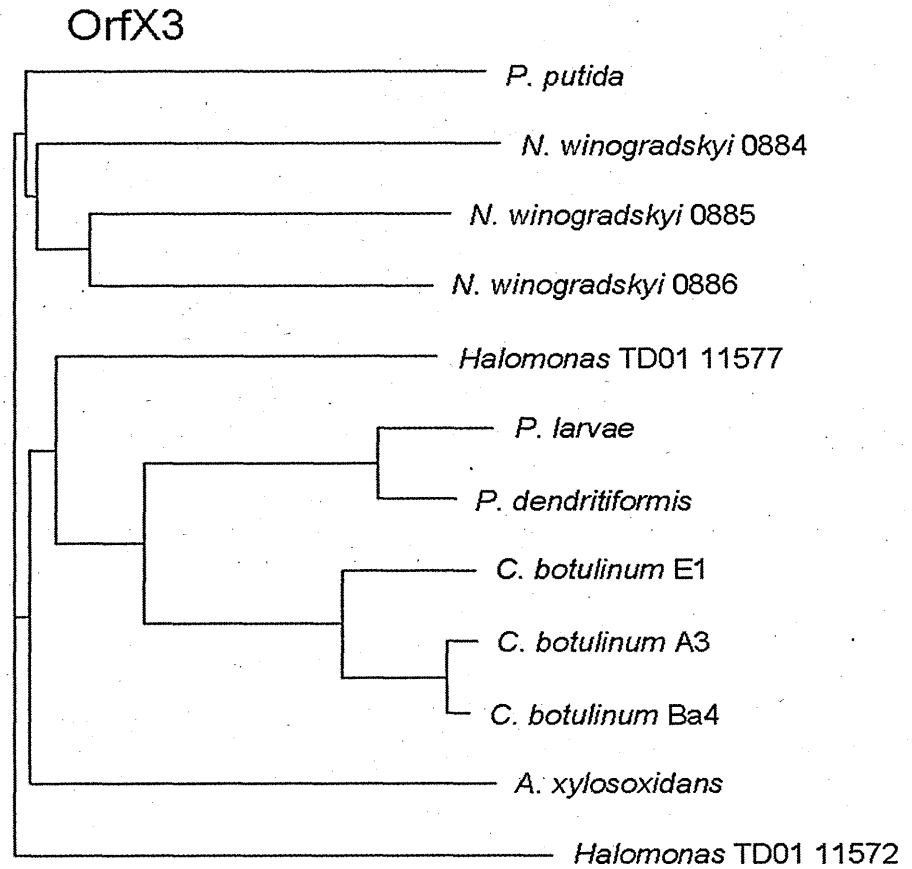


Figure 37: Neighbour joining protein distance tree of OrfX3 sequences from P-47 family encoding species. The *C. botulinum* and *Paenibacillus* sequences cluster together while the other sequences are broadly distributed with no distinct clustering.

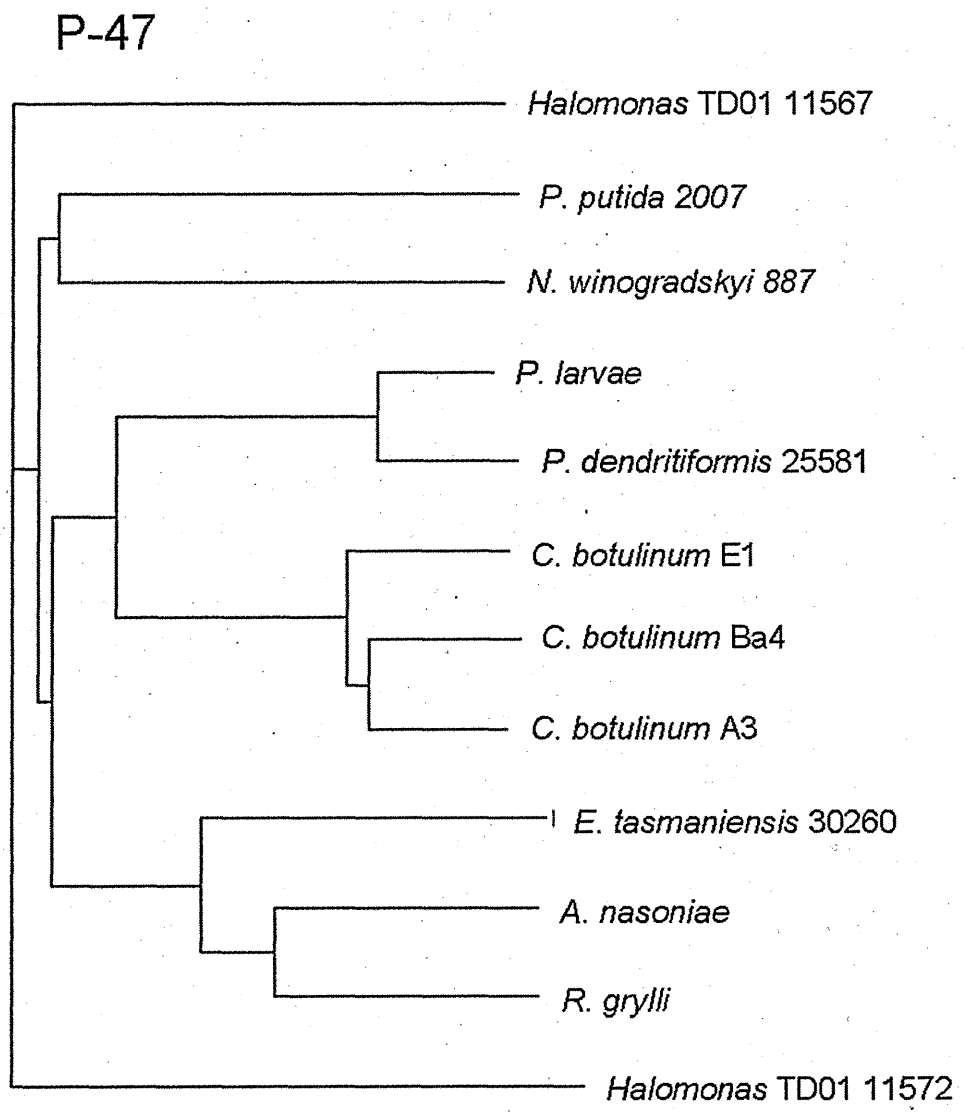


Figure 38: Neighbour joining protein distance tree of P-47 sequences from P-47 family encoding species. The P-47 sequences from *C. botulinum* cluster together with the *Paenibacillus* sequences on an adjoining branch. The sequences from *A. nasoniae, E. tasmaniensis* and *R. grylli* cluster together, similarly to the OrfX2 tree. The *P. putida, Halomonas* TD01 gene GME_11567 and *A. xylosoxidans* sequences cluster individually, distinctly from the other sequences.

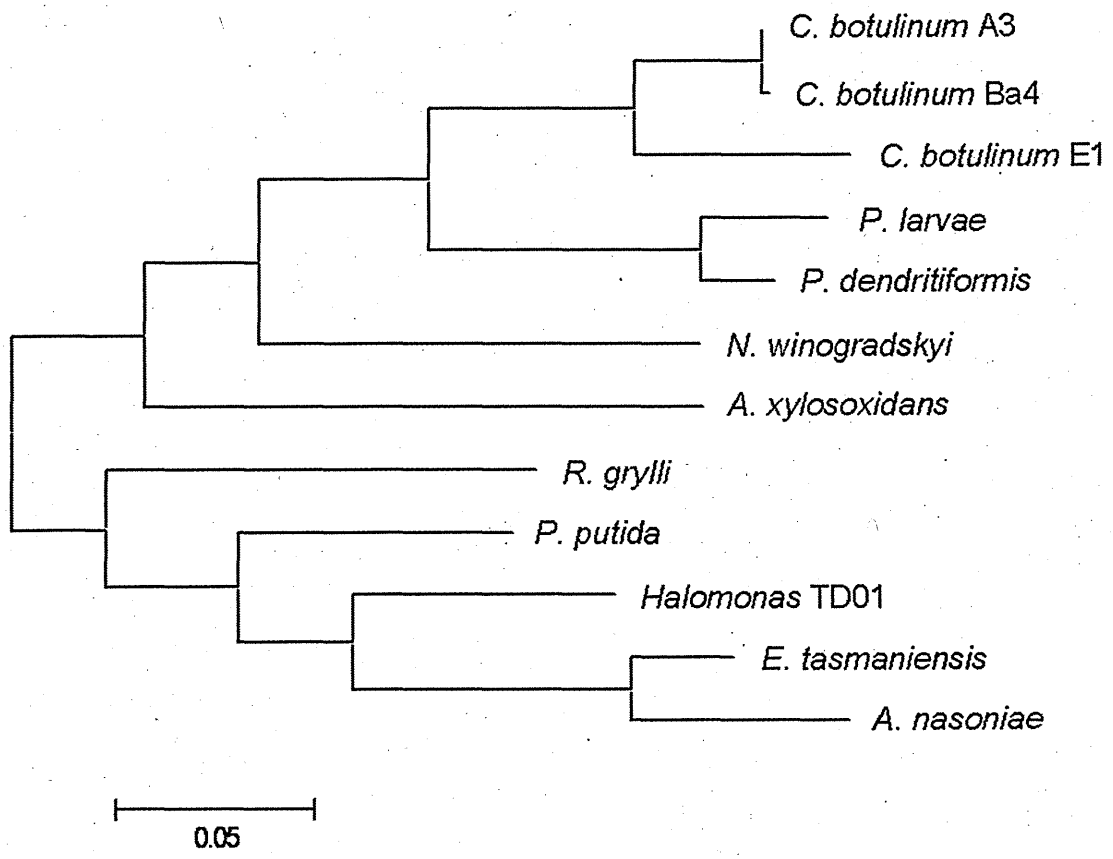


Figure 39: Phylogeny of the P-47 family encoding species constructed from 16S rDNA sequences. The *C. botulinum* group I and group II strains cluster together in independent clusters. The *Paenibacillus* species cluster together on a branch neighbouring the *C. botulinum* strains. *N. winogradskyi* and *A. xylosoxidans* cluster independently of any other species. *R. grylli*, *Halomonas*, *E. tasmaniensis* and *A. nasoniae* form a cluster of gammaproteobacteria.

3.1.3. Prediction of supernatant proteins of *C. botulinum*

The sub-cellular location of the 3550 predicted proteins in the *C. botulinum* A1 ATCC 19397 genome were predicted by five *in silico* tools (CELLO, LocateP, PsortB, SecretomeP and SignalP). The number of proteins predicated as extracellular by each tool and the number of proteins predicated as extracellular by a consensus of 1 to 5 tools was determined (summary Table 27).

There was a range of estimates of the scale of the *C. botulinum* extracellular proteome from 115 proteins estimated by LocateP to 460 by CELLO. There were 11 proteins predicted to be extracellular by all 5 tools. Of these 11 proteins 6 were thermolysin metallopeptidases, involved in protein degradation. Two of the 11 proteins were NlpC/P60 family proteins – this diverse group of proteins includes cell wall peptidases. Another 2 of the 11 proteins are involved in polysaccharide metabolism and the last protein is involved in lipid degradation.

The 113 proteins predicted to be extracellular by four tools included clostripain, the putative toxin activator, eight cell envelope proteins, three protein fate proteins and 73 hypothetical proteins.

There were 235 proteins predicted by three tools to be extracellular including BoNT and NTNH. The tools which did not predict BoNT and NTNH as extracellular was LocateP and SignalP. There were 124 proteins predicted as extracellular by two tools, included in this group were the neurotoxin associated haemagglutinin proteins – the tools which didn't predict these as extracellular were SignalP, LocateP and PsortB. There were 72 proteins predicted as extracellular by one tool.

Table 27: The number of *C. botulinum* A1 19397 proteins predicted to be extracellular by 5 different tools (Cello, locatep, psortb, secretomep and signalp) and the number of *C. botulinum* proteins which were predicted to be extracellular by multiple tools.

Tool	Number of proteins predicted as extracellular
cello	460
locatep	115
psortb	156
secretomep	426
signalp	379

Number of tools	Number of proteins predicted as extracellular
5	11
4	113
3	235
2	124
1	72
Total	556

3.1.4. Summary of findings of in silico investigation

- There are multiple coding sequences with similarity to P-47 family proteins in nine non-*C. botulinum* species
- In four of those nine species, the P-47 family clusters are co-localised with putative toxin genes.
- This evidence supports the hypothesis that the P-47 family cluster plays a role in the toxicity of BoNT.
- Sub-cellular location prediction tools gave varied results, with significant overlap. Comparison of the predictions of the different tools against experimentally identified supernatant proteins will allow informed discussion of their relative merits.

3.2. Proteomic investigation of *C. botulinum* to establish protein profiles associated with toxin producing strains

The mechanism of action of botulinum neurotoxin (BoNT) at the human neuromuscular junction is well understood. However, the variety of botulism disease types and severities suggests that there may be a complex interaction of BoNT with other *C. botulinum* proteins, influencing disease outcome. Previous reports (Fujinaga, 2010) have implicated proteins encoded by genes co-located and co-expressed with the *bont* gene in the toxicity of BoNT via the oral route. This suggests that variation in these proteins may explain variation in disease severity. The proteins produced by such genes (the neurotoxin associated proteins) form large complexes with BoNT that range in size between 290-900 kDa. The role of these neurotoxin-associated proteins in the botulism disease process is not fully understood. It has been reported that some of the proteins contribute to toxicity via the oral route by protecting the toxin from degradation in the gut and also that they have a role in translocation of the toxin across the gut epithelium (Simpson, 2004; Sugawara et al., 2010).

In this study, the toxin and associated proteins present in culture supernatants of *C. botulinum* reference strains and isolates from clinical cases of food, wound and infant botulism were investigated. Initial experiments to establish the relationship between the growth phase of *C. botulinum* and the amount of extracellular protein were performed. The concentration of BoNT at different growth stages was assessed using an endopeptidase assay. Based on the results of these experiments, a broader characterisation of the supernatant proteome was undertaken at two time-points using LC-MS/MS.

In this chapter the investigation of the presence of the toxin complex, and other uncharacterised pathogenicity factors in the extracellular proteome of *C. botulinum* is reported.

3.2.1. C. botulinum and C. sporogenes growth curves

Growth curves were prepared using three reference strains. Two of these were *C. botulinum* strains: *C. botulinum* A1 ATCC 19397 and *C. botulinum* B NCTC 7273; the other was *C. sporogenes* NCTC 275 which was used as a known non-toxigenic control to optimise procedures before the investigation of toxigenic strains. Growth curves for each strain were determined over a 96 h period.

Growth curves produced for all three strains showed similar properties (Figure 40). *C. sporogenes* NCTC 275 OD_{600} increased 6.9 fold between 6 and 12 h but then decreased to OD_{600} 1.14 at 48 h and plateaued until 96 h when the experiment was terminated. *C. botulinum* A1 ATCC 19397 showed a 9.8 fold increase in OD_{600} between 4 h and 12 h, peaking at an OD_{600} of 3.0, after which it decreased gradually to 0.98 at 48 h, holding at this OD_{600} until 96 h. Similarly, *C. botulinum* NCTC 7273 underwent a 10.6 fold increase in OD_{600} between 4-11 h, peaking at 2.7 before gradually decreasing to 0.79 at 48 h. OD_{600} plateaued between 48 and 96 h.

Cell count was determined using a hemocytometer for one of the strains, ATCC 19397, at the same time-points used to measure the OD_{600} growth curve (Figure 41). The concentration of cells increased approximately 200 fold between 4 h and 14 h, reaching a peak of 1.78×10^9 cells per ml at 14 h, representing the end of logarithmic growth.

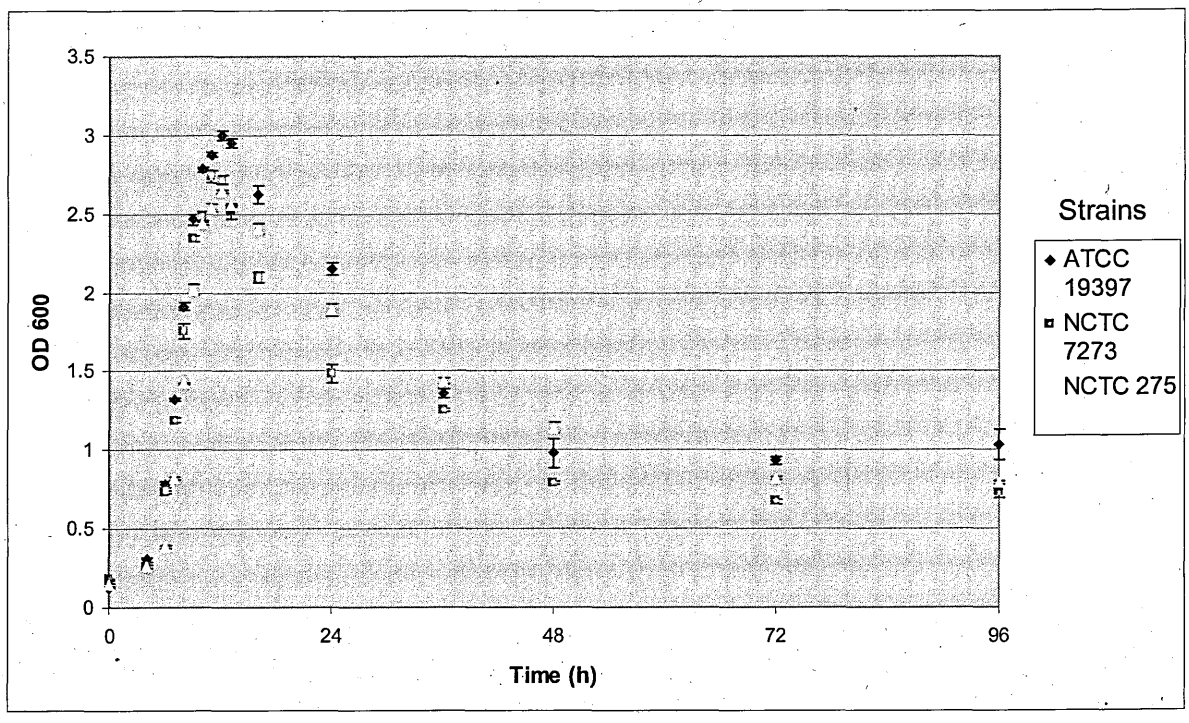


Figure 40: Growth curve of *C. botulinum* ATCC 19397, NCTC 7273 and *C. sporogenes* NCTC 275 under anaerobic incubation in TPGY at 35°C, as measured by OD_{600}.

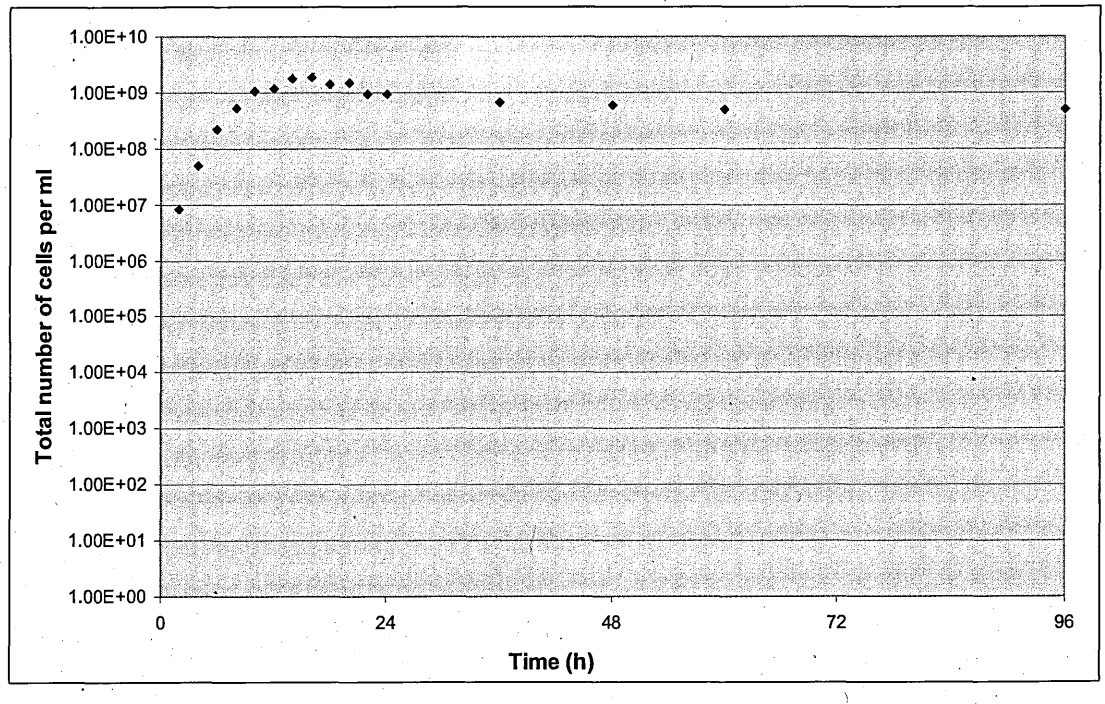


Figure 41: Total cell count of an ATCC 19397 culture grown anaerobically in TPGY at 35°C. Note logarithmic scale.

3.2.2. Optimisation of protein precipitation from *C. botulinum* culture supernatant

3.2.2.1. Comparison of acetone and trichloroacetic acid precipitation

Protein precipitation was optimised using culture supernatant from the non-toxigenic strain *C. sporogenes* NCTC 275. Two different precipitants were used: acetone and trichloroacetic acid (TCA) (Schwarz et al., 2007) and the total amount of protein precipitated and the range of proteins precipitated was compared. The range of proteins precipitated was assessed using SDS-PAGE. Having determined which precipitant was more suitable, the concentration of precipitant was optimised.

After precipitation, proteins were re-suspended in a buffer containing 1% SDS. This is above the level of SDS that is compatible with the widely used Bradford assay (0.125%), therefore the SDS compatible bicinchoninic acid (BCA) assay was used for quantification of precipitated proteins.

Acetone precipitated proteins could not be accurately quantified by the BCA assay, due to an unknown interfering substance. Densitometric comparison of the TCA precipitant with the acetone precipitant indicated that acetone precipitation yielded a lower concentration of protein (Figure 42). There was substantially less protein in 6.5 µl of acetone precipitant than there was in 2.2 µl of TCA precipitant. In addition to lower total yield, there was also a total of 40 protein bands from TCA precipitation compared with 16 protein bands from acetone precipitation detected by densitometric comparison (Figure 42). As TCA precipitated a better yield of a

wider range of protein bands, and no accurate quantification of acetone precipitated samples could be achieved, TCA was used for subsequent protein precipitation.

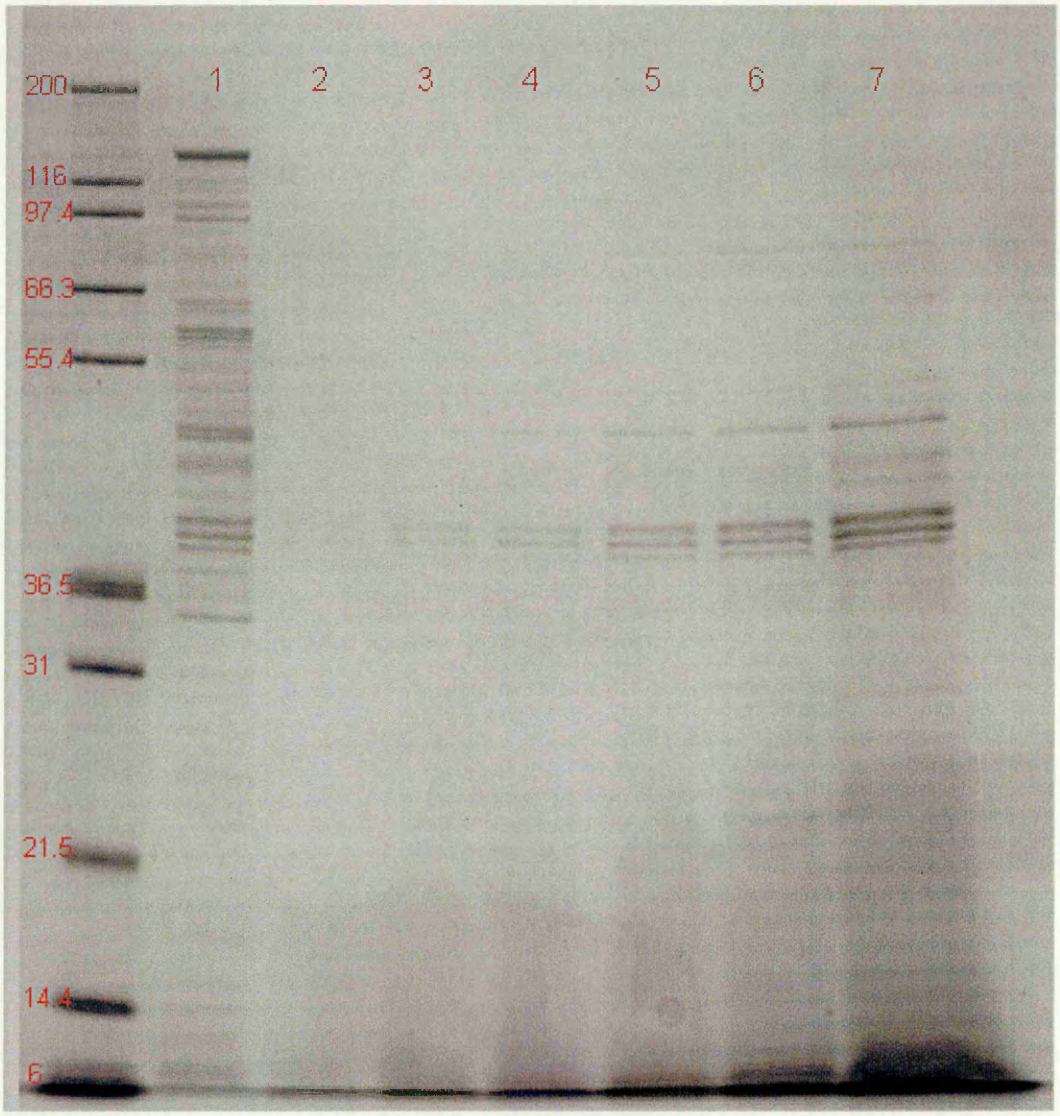


Figure 42: *C. sporogenes* NCTC 275 supernatant protein precipitated with TCA and acetone. Lane 1: 5 µg TCA (2.2 µl) precipitated protein. Lanes 2-7: 0.1, 0.5, 1, 2, 3 and 6.5 µl of acetone precipitated protein.

3.2.2.2. Optimisation of TCA concentration

Protein precipitation was further optimised by investigating the amount and range of protein precipitated from culture supernatant by three concentrations of TCA. Culture supernatant from 16 h and 96 h of growth was examined to ensure that culture age had no impact on protein precipitation. TCA was used at 5%, 10% and 20% (v/v), to precipitate proteins from 16h and 96 h cultures of *C. sporogenes* NCTC 275. A TCA concentration of 5% gave a 25±7% higher yield of protein at 24 h and a 17±6% higher yield of protein at 96 h than either 10% or 20% TCA (Figure 43). All TCA concentrations gave a similar protein profile when compared using 1DGE (Figure 44) and therefore 5% TCA was used in future experiments to precipitate *C. botulinum* supernatant proteins.

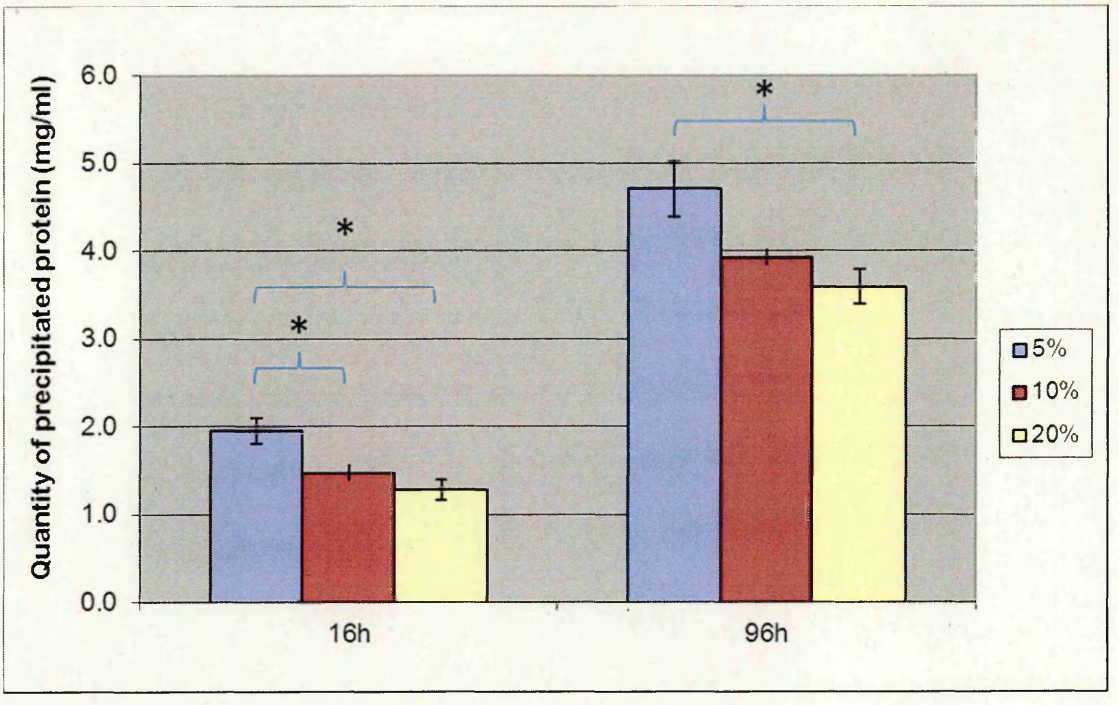


Figure 43: Average amount (n = 3) of protein precipitated from *C. sporogenes* NCTC 275 supernatant from 16h and 96 h cultures using 5, 10 and 20% TCA. Protein concentration measured by BCA assay. Asterisks denote t-test p-value of < 0.05.

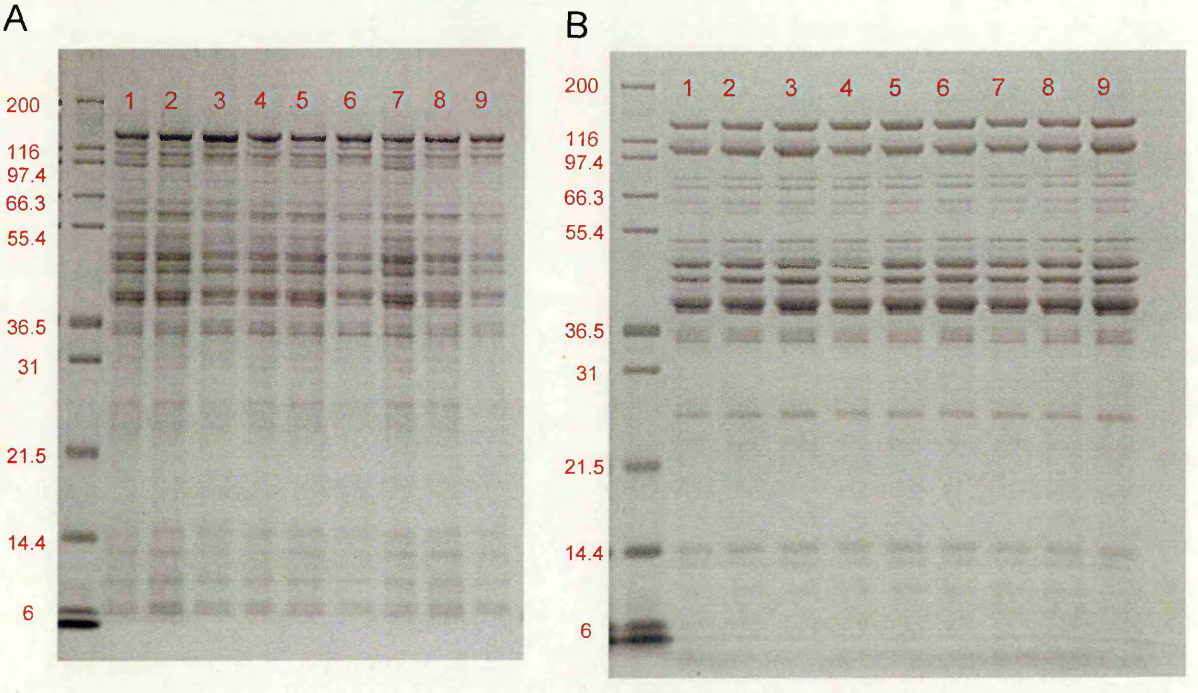


Figure 44: SDS-PAGE of protein precipitated from *C. sporogenes* culture supernatants from cultures grown at (A) 16 h amd (B) 96 h obtained using TCA at 5% (lanes 1-3), 10% (lanes 4-6) and 20% TCA (lanes 7-9).

3.2.2.3. Comparison of protein concentration of total supernatant with protein concentration of precipitated protein

The amount of protein that was precipitated from the supernatant of a *C. sporogenes* culture was compared to the concentration of total protein in the supernatant. Two methods of protein quantification, the Bradford and BCA assays, were assessed for their ability to accurately quantify the protein in the total supernatant. The BCA assay was unsuitable for measuring the protein concentration of the total supernatant due to the presence of an unknown interfering agent in the sample. This interference was not removed by dilution or by filtering the sample though a 3 kDa molecular weight cut off filter. Therefore the Bradford assay was used to measure the concentration of the total supernatant protein and the BCA assay was used to measure the concentration of protein in the precipitant (as the Bradford assay is not compatible with the detergents present in the re-suspended protein precipitant).

Initial results showed that the amount of TCA precipitated protein was approximately 10% of the protein content of the total supernatant (data not shown). However, when 5 µg of each was analysed by SDS-PAGE (Figure 45), lower concentrations of protein were observed for the total supernatant, indicating that the Bradford procedure had overestimated the concentration of protein in the supernatant. This overestimation was determined to be 3.15x by densitometric analysis of the SDS-PAGE gel (Figure 45). When the values obtained by the Bradford assay for total supernatant protein content were, therefore, divided by 3.15, TCA precipitation was estimated to have extracted >75% of the total protein in the supernatant (Figure 46). Although the supernatant proteins are under-

loaded the SDS-PAGE separation shows that the precipitated protein and the supernatant protein profiles are very similar (Figure 45).

A comparison of the total supernatant protein concentration and the precipitated protein concentration along a time course was carried out. The amount of protein in the total supernatant increased 6.1 fold between 0 and 12 h whilst that in the precipitant increased 4.7 fold. There was a slower, gradual increase in supernatant protein and precipitated protein between 12 and 96 h, which peaked at 5.1 and 3.9 mg/ml respectively.

Following optimisation of the protein precipitation method using *C. sporogenes* NCTC 275, a *C. botulinum* ATCC 19397 culture time course was carried out in triplicate and supernatant protein was precipitated with 5% TCA at 8 time points (Figure 47). There was a gradual 4-fold increase in the concentration of precipitated protein between 0 and 72 h. Between 72-96 h there was a 25% decrease in the protein precipitant.

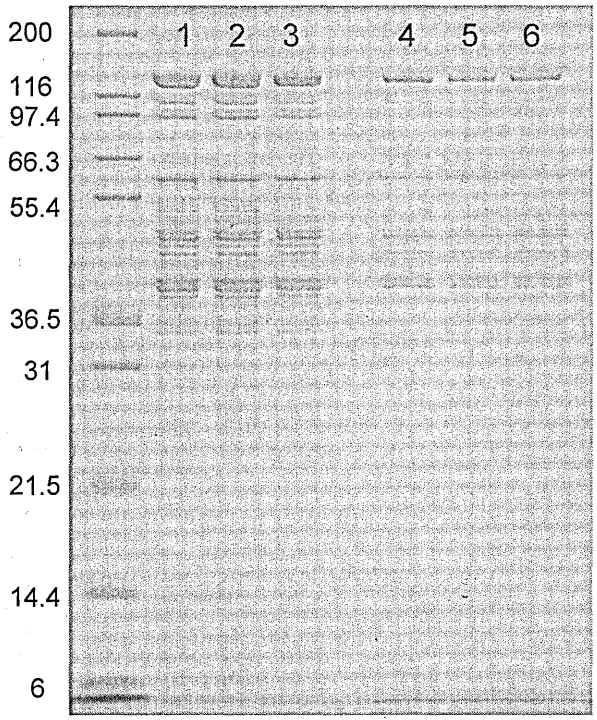


Figure 45: SDS-PAGE gel of precipitated supernatant protein and original supernatant protein from 24 h culture of *C. sporogenes* NCTC 275. Lanes 1-3 contain biological replicates of protein precipitated from the supernatant using 5% TCA and lanes 4-6 contain total supernatant protein from the same cultures (no precipitation). Lanes 4-6 are under-loaded compared with lanes 1-3 as accurate quantification of total supernatant protein was not possible.

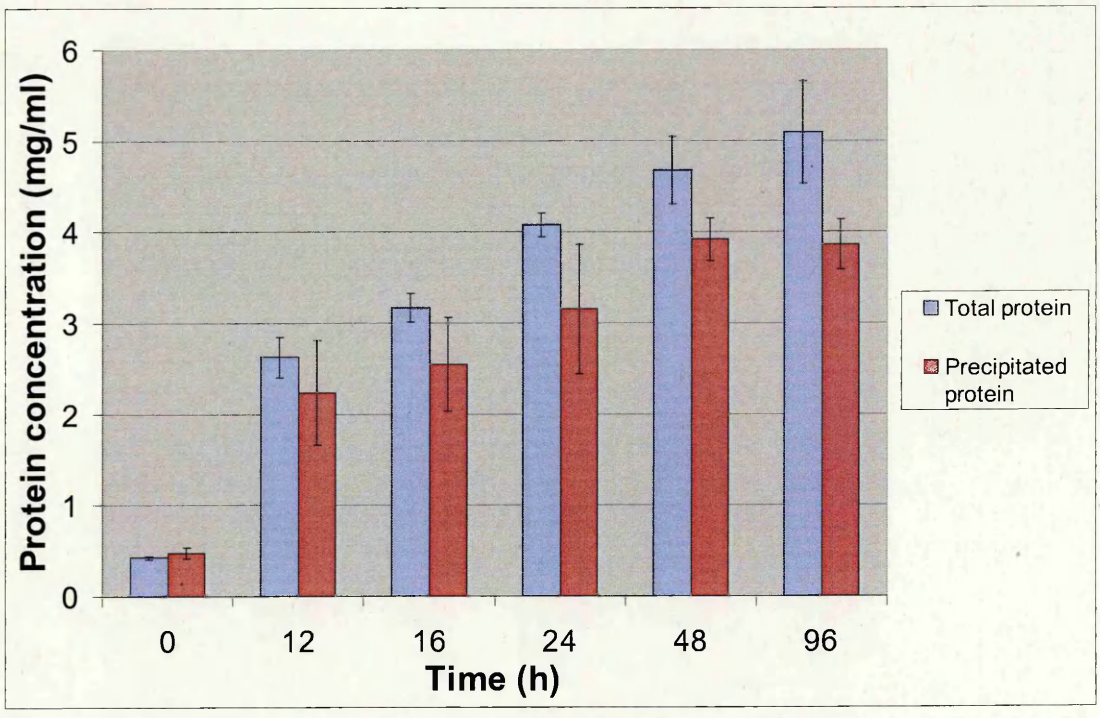


Figure 46: Average (n = 3) protein concentration of *C. sporogenes* NCTC 275 total supernatant (adjusted by densitometric analysis) and TCA precipitant between 0-96 h.

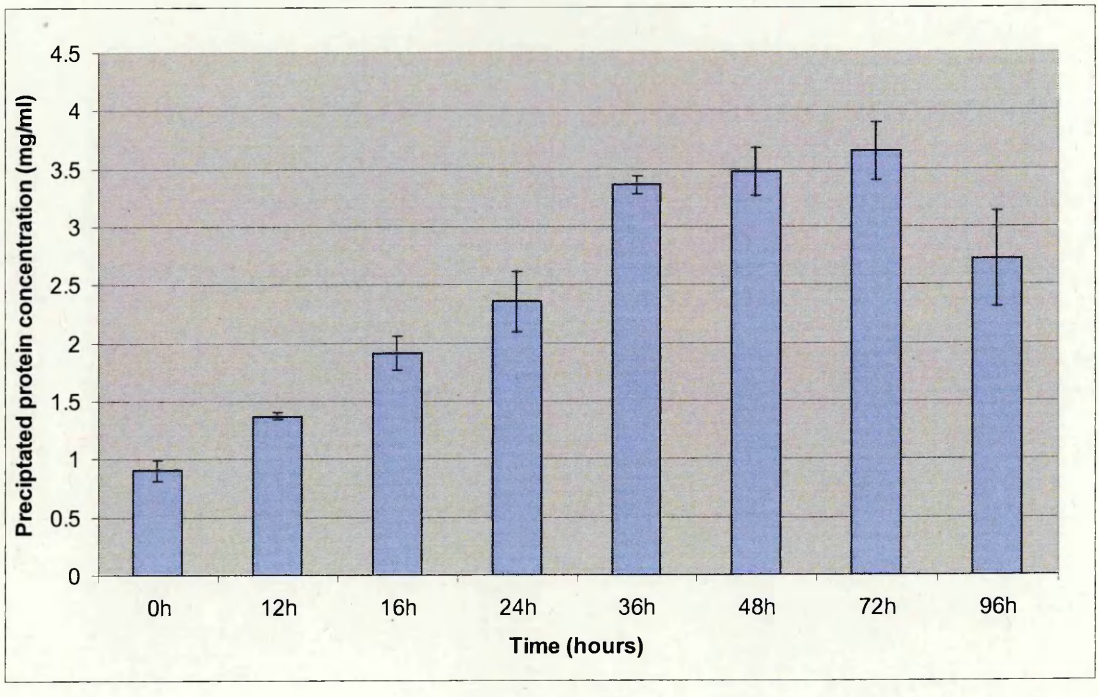


Figure 47: Average (n = 3) concentration of protein precipitated from the supernatant of *C. botulinum* A1 ATCC 19397 between 0-96 h.

3.2.3. Determination of toxin concentration in the culture supernatant using endopeptidase assay

The concentration of active BoNT/A in the culture supernatant was determined using an endopeptidase activity assay (Figure 48) (as described in materials and methods 1.3.1). This assay involves incubating the culture supernatant with synthetic SNAP-$25_{137-206}$, the fragment of the BoNT/A substrate that contains the target site. An HRP-conjugated antibody, specific to the $SNAP25_{190-197}$ octapeptide epitope of the BoNT/A cleaved substrate, allowed quantification of the BoNT/A activity. The results are expressed as MLD_{50} values (lethal dose for 50% of inoculated mice per ml) - MLD_{50} equivalents could be calculated as the culture supernatant was compared against a dilution series of toxin standards of known concentration which had been previously established (Sesardic et al., 2003; Jones et al., 2008). There were 31464 MLD_{50}/ml at 0 h (Figure 48), this is likely to be the result of residual toxin present in the 1 ml suspension of ATCC 19397 overnight culture used to inoculate the culture. The supernatant toxin concentration rapidly increased 15.8-fold, reaching 497337 MLD_{50}/ml by 24 h. The concentration of toxin then gradually declined between 24 h and 96 h, decreasing to 32830 MLD_{50}/ml.

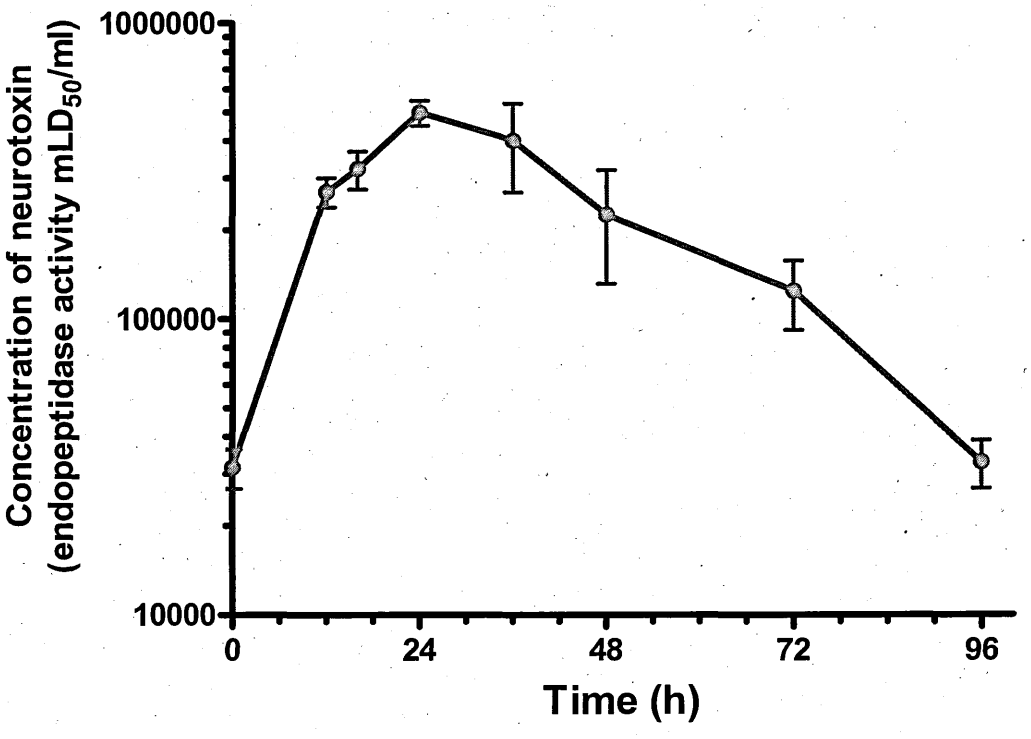


Figure 48: BoNT/A activity of *C. botulinum* A1 NCTC 19397 culture supernatant incubated anaerobically at 35°C. Eight samples were taken between 0-96 h (average of three biological replicates) and BoNT/A activity measured by endopeptidase assay.

3.2.4. Detection and identification of botulinum neurotoxin and other proteins in *C. botulinum* culture supernatant by LC-MS/MS

The whole culture supernatant proteome, including botulinum toxin and the associated non-toxic proteins were analysed using an approach involving initial separation by SDS-PAGE followed by LC-MS/MS. As the toxin and its associated proteins are co-transcribed the culture supernatant was investigated at 24 h, the time at which the endopeptidase assay indicated peak toxin concentration (Figure 48). The supernatant proteome was also investigated at 96 h.

In terms of data analysis, all samples were first analysed against a database including all NCBI non-redundant protein sequences to ensure that the majority of identifications were against *C. botulinum* strains. If this was the case then the proteomic data was compared against a single appropriate reference genome.

3.2.4.1. Analysis of *C. botulinum* A1 ATCC 19397 culture supernatant after 24 and 96 h culture

Supernatant proteins from 24 and 96 hours of broth culture growth (n = 3) were separated and analysed by LC-MS/MS. There were a total of 198 proteins detected and identified at 24 h and 144 proteins identified at 96 h. These proteins were compared using local BLAST which identified 63 proteins unique to culture supernatant at 24 h, a large stable proteome of 135 proteins identified at both time points and only nine proteins unique to culture supernatants grown for 96 h (a complete listing is given in Appendix tables 2-4).

LC-MS/MS analysis revealed that BoNT, NTNH, HA70, HA33, HA17 and clostripain (a protease implicated in *C. botulinum* pathogenesis) were all expressed at both 24 h and 96 h. The regulator, botR was not identified LC-MS/MS detection of toxin in the supernatant at both time points confirmed the results determined by the endopeptidase assay.

Identified proteins were assigned to one of 21 functional groups (e.g. energy metabolism, cell membrane, etc) using protein annotations from the online resource, Pathema (http://pathema.jcvi.org/cgi-bin/Clostridium/PathemaHomePage.cgi). The number of proteins belonging to each functional group that were unique to 24 h, unique to 96 h or present at both time points were compared (Table 28).

Only four functional groups had all their proteins identified at both time-points – pathogenesis, signal transduction, DNA-metabolism and regulatory function associated proteins. Pathogenesis-associated proteins were the largest of these four groups, with 6 proteins (BoNT, NTNH, HA70, HA33, HA17 and Clostripain). There were two signal transduction proteins (both involved with phosphotransferase systems, Uniprot IDs = A7FVX2, A7FVJ3), one regulatory function protein (DNA binding protein, A7FSQ9) and one DNA metabolism protein (single stranded DNA binding protein, A7FZH0) identified at both time-points.

There was a general trend that more proteins were detected in each functional group at 24 h than 96 h. In keeping with this trend there were 20 more proteins of unknown or hypothetical function, nine more energy metabolism proteins (including thioredoxin family protein, A7FTD9 & arginine deiminase, A7FQ94) and six more protein synthesis associated proteins (including 30S and 50S ribosomal

proteins, A7FQ40 & A7FZ57) present in the culture supernatant at 24 h than at 96 h. There were also three more transport and binding proteins (three ATP binding cassette transporters including a molybdate transporter, A7FU00) and three more cell envelope construction and maintenance proteins (including a lipoprotein, A7FU16) present at 24 h than 96 h. There were additional proteins involved in the biosynthesis of amino acids and the biosynthesis of cofactors, prosthetic groups and carriers (including thiamine biosynthesis protein, A7FXG9) at 24 h than 96 h. The only two functional groups represented at 24 h that were not identified at 96 h were adaptations to atypical conditions (major cold shock protein *CspA*, A7FTR1) and mobile and extrachromosomal element functions (phage capsid protein, A7FWD9) that have one protein each at 24 h (Table 28).

Detoxification was the only functional group with more proteins at 96 h than at 24 h; 3 proteins at 96 h versus 2 at 24 h. There were the same number of pathogenesis-associated proteins present at both 24 and 96 h (Table 28).

Table 28: *C. botulinum* A1 ATCC 19397 supernatant proteins present at 24 h and/or 96 h incubation according to functional group.

Functional role	24 h only	24 and 96 h	96 h only
Adaptations to atypical conditions	1		
Amino acid biosynthesis	4	2	
Biosynthesis of cofactors, prosthetic groups, and carriers	3	2	1
Cell envelope	3	6	
Central intermediary metabolism	3	2	
Chemotaxis and motility	1	4	
Detoxification		2	1
DNA metabolism		1	
Energy metabolism	10	43	1
Fatty acid and phospholipid metabolism	1	5	
Hypothetical and unknown proteins	20	22	4
Mobile and extrachromosomal element functions	1		
Pathogenesis		6	
Protein fate	4	19	
Protein synthesis	6	9	1
Purines, pyrimidines, nucleosides, and nucleotides	1	4	
Regulatory functions		1	
Signal transduction		2	
Sporulation and germination	1		1
Transcription	1	4	
Transport and binding proteins	3	1	
Total	**63**	**135**	**9**

3.2.4.2. Comparison of *C. botulinum* A1 ATCC 19397 supernatant proteins with the proteins in the MvirDB virulence database

The MvirDB (http://predictioncenter.llnl.gov/) is a database of toxins, virulence factors and antibiotic resistance associated proteins. BLAST was used to search for sequence similarity between the 207 proteins identified in the supernatant proteome of *C. botulinum* A1 ATCC 19397 and the 64000 proteins in the MvirDB in order to identify potential novel virulence factors. There were two ATCC 19397 proteins (not previously annotated as involved in pathogenesis) that showed significant homology with virulence proteins in the MvirDB (for more details see methods section).

The first *C. botulinum* A1 ATCC 19397 protein to show homology to a protein in MvirDB was a thermolysin metallopeptidase (A7FTX0) that has identical functional moieties in a very similar architecture to the *C. perfringens* lambda toxin (Figure 49). The two proteins have 30.5% amino acid identity. Both proteins have M4 metallopeptidase domains (IPR001570) spanning amino acid residues 264-580 of the 580 amino acid *C. botulinum* protein and residues 249-553 of the 553 amino acid *C. perfringens* protein. M4 metallopeptidase domains are typical of thermolysins which are thermo-stable, secreted proteases. Both proteins also have a pro-peptide domain, PepSY (IPR025711), at residues 166-217 of the *C. botulinum* protein and residues 151-201 of the *C. perfringens* protein. This pro-peptide prevents premature inactivation of the metalloprotease. The *C. perfringens* lambda toxin is involved in the proteolytic cleavage and activation of the *C. perfringens* epsilon proto-toxin (Jin et al., 1996).

The second ATCC 19397 protein to show significant sequence similarity to a protein in the MvirDB was an ATP-dependent Clp protease (GI 153933152). The ATCC 19397 Clp protease showed 66% amino acid identity across 87% of the protein length to a *Listeria monocytogenes* ATP-dependent Clp protease (GI 16804506) (Figure 50). This is an extracellular protease implicated in *L. monocytogenes* interaction with the host including macrophage escape and activation of listeriolysin (Gaillot et al., , 2000). The *C. botulinum* and *L. monocytogenes* proteins share the ATP-dependent protease ClpP region (IPR001907) which spans the entire length of both proteins. The ClpP active site is in the same location in both proteins, consisting of two domains between residues 95-135 of the 194 amino acid *C. botulinum* protein and residues 96-136 of the 198 amino acid *L. monocytogenes* protein.

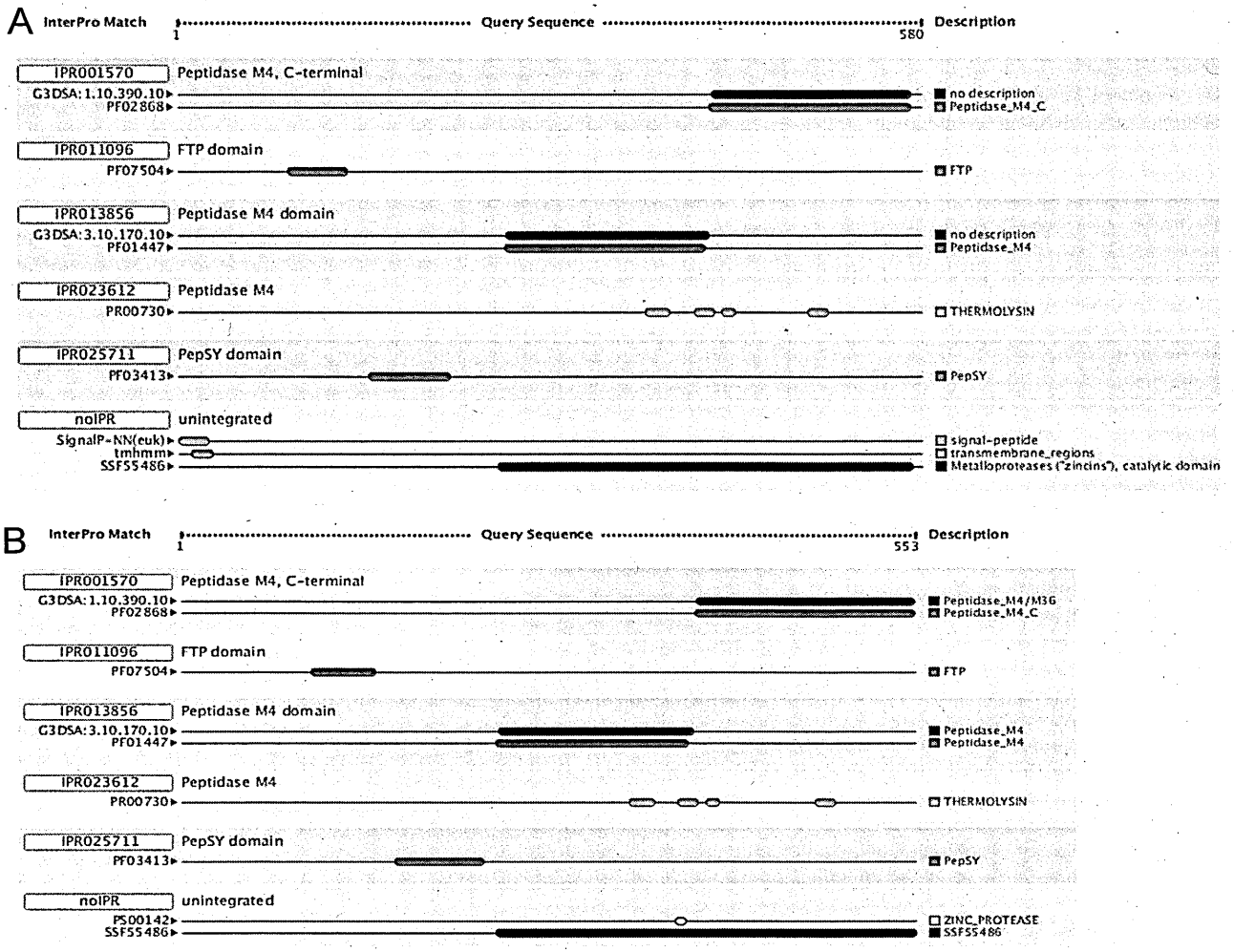


Figure 49: Comparison of the functional moieties of (A) *C. botulinum* A1 19397 thermolysin metallopeptidase (GI 153931929) and (B) *C. perfringens* lambda toxin as determined by Interproscan (http://www.ebi.ac.uk/Tools/pfa/iprscan/).

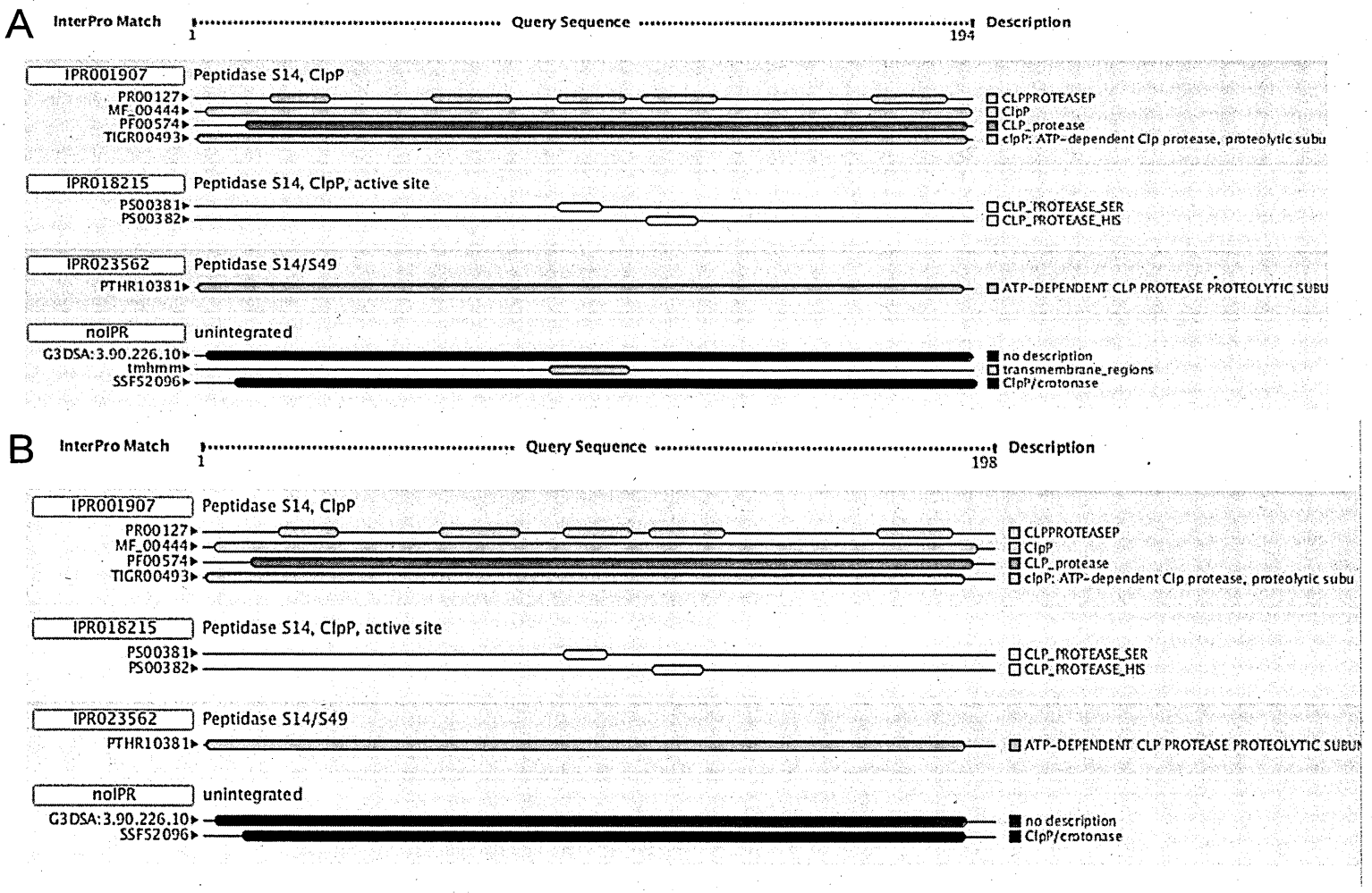


Figure 50: Comparison of the functional moieties of (A) *C. botulinum* A1 19397 ATP dependent Clp protease (GI 153933152) and (B) *L. monocytogenes* ATP-dependent Clp protease as determined by Interproscan.

3.2.4.3. Analysis of the supernatant proteome of *C. botulinum* B NCTC 7273 using LC-MS/MS

C. botulinum B NCTC 7273 supernatant proteins were extracted 24 h after inoculation (n = 3) and analysed using the 1DGE LC-MS/MS workflow. There were 118 proteins which were detected and identified in all three biological replicates (a complete list is given in Appendix Table 5). Identified proteins were assigned to one of 21 functional groups (e.g. energy metabolism, cell membrane, etc) using protein annotations from Pathema.

The toxin complex proteins - BoNT/B (Uniprot ID = B1INP5), NTNH (B1INP6), HA33 (B1INP8), HA17 (B1INP9) and HA70 (B1INQ0), were all detected in the supernatant of NCTC 7273 at 24 h. Clostripain (B1IMT3), an enzyme potentially involved in the proteolytic activation of the toxin was also identified. Other proteins identified in the supernatant of NCTC 7273 included three amino acid biosynthesis proteins, including ornithine carbamoyltransferase (B1IJJ6). There were three cell envelope proteins identified, including a myosin-cross-reactive antigen family protein (B1IFN3). There were four proteins identified which have a role in chemotaxis or motility including a flagellin protein (B1IK59) and a CheA chemotaxis protein (B1IKG0). Two sporulation and germination proteins were identified – both stage V, one G protein (B1IH03) and one S protein (B1II33). The protein functional group that contained the largest number of proteins identified in the supernatant of NCTC 7273 was energy metabolism, with 35 proteins. Of these, eight were involved in amino acid energy metabolism, including two arginine deiminases (B1ID40), eight were involved with fermentation, including alcohol dehydrogenase (B1IKP6) and six were involved in glycolysis/gluconeogenesis including a glucose-6-phosphate isomerise (B1ING9). There were two proteins involved in fatty acid and phospholipid metabolism, including an acyl-CoA

dehydrogenase which is involved in fatty acid degradation. There were 16 proteins which were hypothetical or of unknown function. The second most represented group of proteins identified in the NCTC 7273 supernatant were involved in protein fate, 20 of these proteins were identified. Of these 20, 14 were involved with the degradation of proteins including a collagenase (B1ILP0) and numerous peptidases including peptidase T (B1IEF5). Other proteins identified that have a role in protein fate included four protein folding and stabilisation proteins including chaperonin GroEL (B1IFD5). There were nine protein synthesis proteins, four of which were tRNA aminoacylation proteins including methionyl-tRNA synthetase (B1ID49), three translation factors including translation elongation factor Tu (B1IGF6). There were two signal transduction proteins, both involved in phosphotransferase systems including a phosphocarrier protein HPr (B1II28). There were four transport and binding proteins identified including an amino acid transport protein (B1IKD4).

3.2.4.4. Comparison of the supernatant proteome of *C. botulinum* B NCTC 7273 with the virulence database MVirDB

The proteins identified in all replicates of the *C. botulinum* B NCTC 7273 supernatant were compared against the MvirDB virulence database. One *C. botulinum* B NCTC 7273 protein that had a high quality match (see materials and methods) for a protein present in MvirDB was a thermolysin metallopeptidase (B1IKQ2). This protein had a similar architecture to both the *C. perfringens* lambda toxin and the *C. botulinum* A1 ATCC 19397 thermolysin metallopeptidase that showed homology to the lambda toxin (Figure 51). Both proteins have M4 metallopeptidase domains (IPR001570) spanning amino acid residues 264-579 of the 579 amino acid *C. botulinum* B NCTC 7273 protein and residues 249-553 of

the 553 amino acid *C. perfringens* lambda toxin. Both proteins have a pro-peptide domain, PepSY (IPR025711), at residues 164-216 of the *C. botulinum* protein and residues 151-201 of the *C. perfringens* protein. This pro-peptide prevents premature inactivation of the metalloprotease. The *C. perfringens* lambda toxin is involved in the proteolytic cleavage and activation of the *C. perfringens* epsilon proto-toxin (Jin et al., 1996). The *C. botulinum* B NCTC 7273 thermolysin and the *C. botulinum* A1 ATCC 19397 protein shared 86.9% amino acid identity while the *C. botulinum* B NCTC 7273 thermolysin metallopeptidase has 30.4% identity with the *C. perfringens* lambda toxin.

Another *C. botulinum* B NCTC 7273 protein highlighted as a potential virulence factor by comparison with MvirDB is a collagenase protein (B1ILP0) that shows 38.9% amino acid identity and very similar protein architecture with a *C. perfringens* collagenase. Both proteins have a domain indicative of extracellular proteins (IPR022409) and multiple domains typical of a collagenase (IPR013661, IPR007280 and IPR002169), in the same order along the length of the protein.

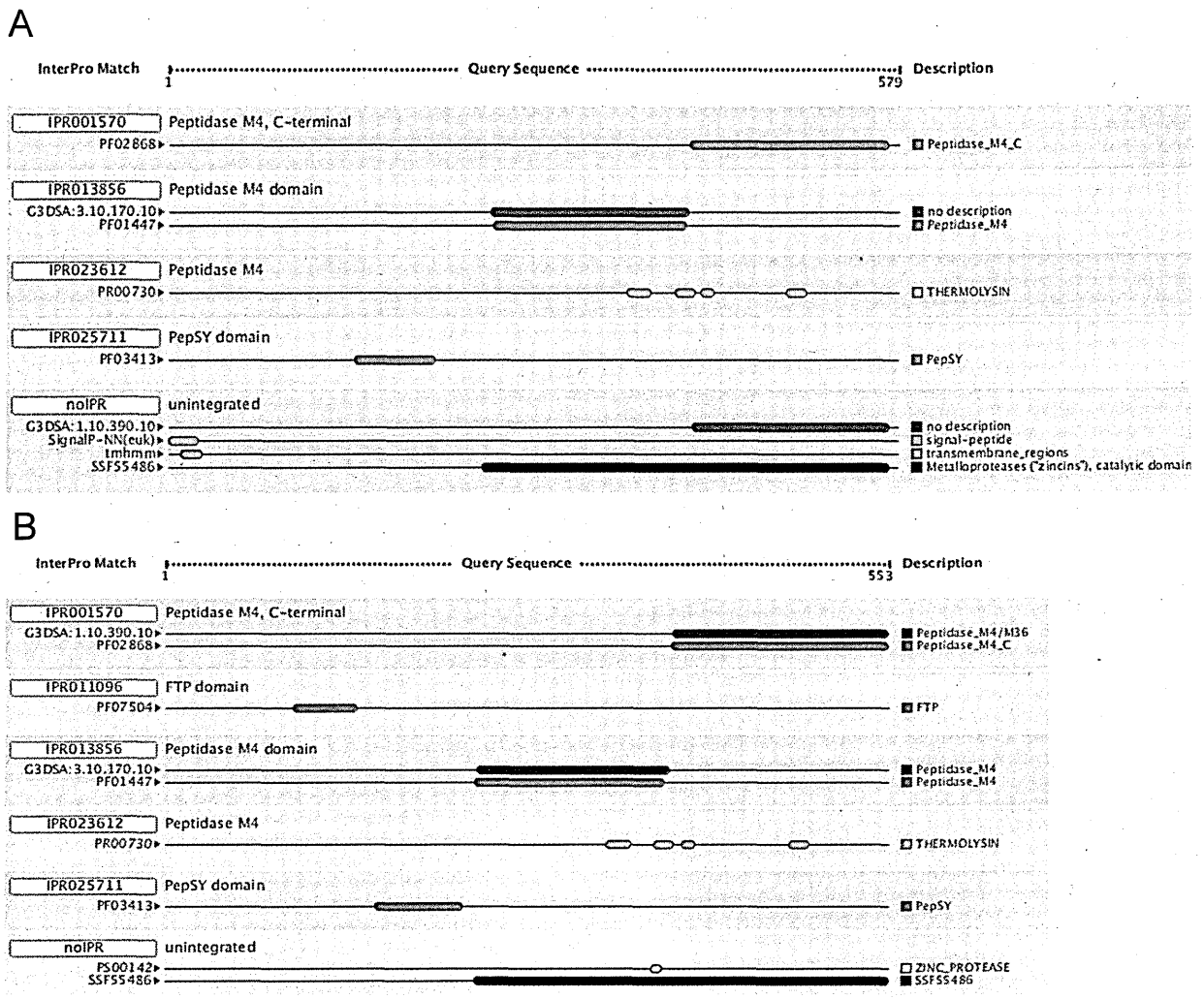


Figure 51: Comparison of the functional moieties of (A) *C. botulinum* B NCTC 7273 thermolysin metallopeptidase (B1IKQ2) and (B) *C. perfringens* lambda toxin as determined by Interproscan

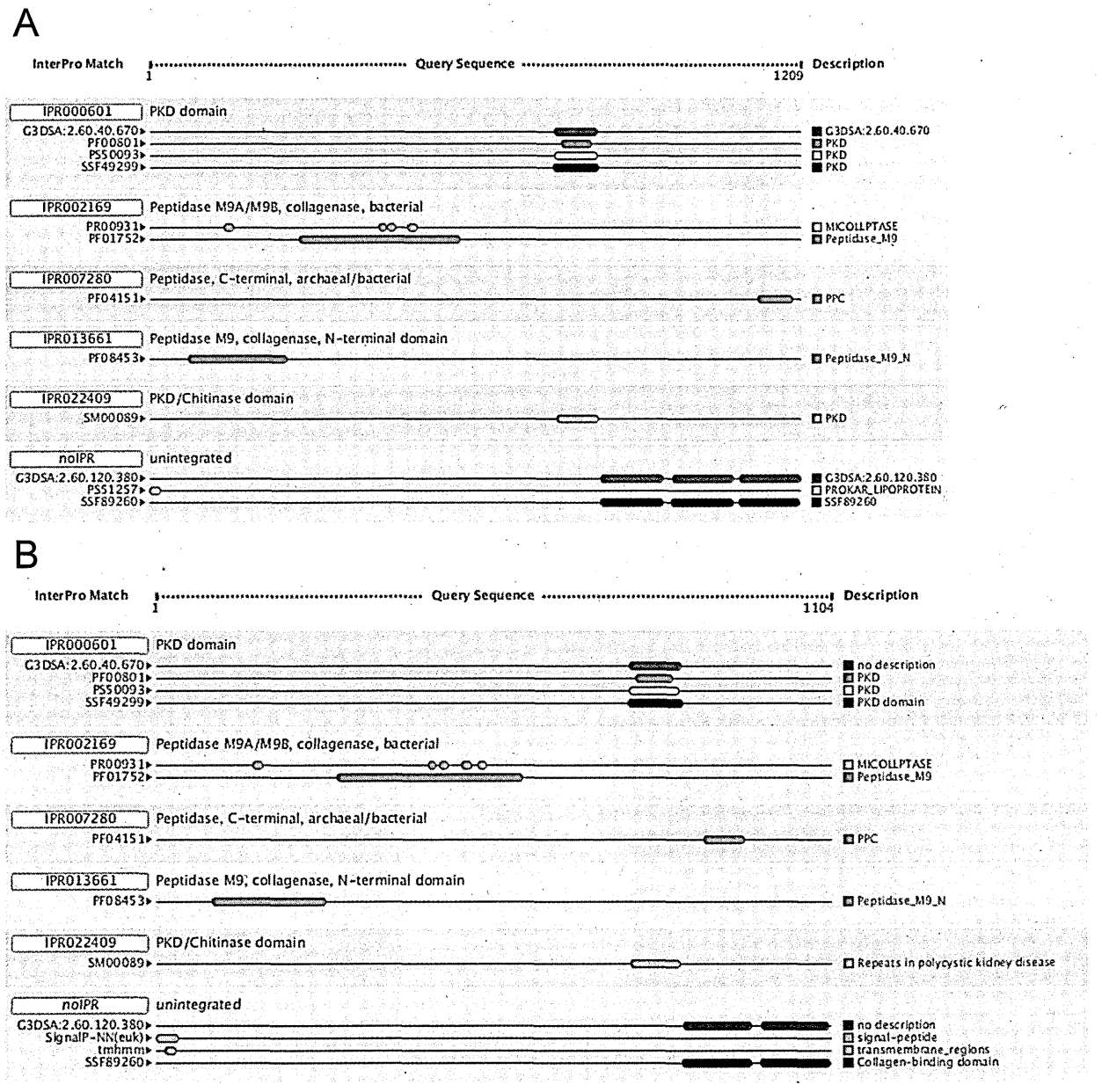


Figure 52: Comparison of the functional moieties of (A) *C. botulinum* B NCTC 7273 collagenase (B1ILP0) and (B) *C. perfringens* collagenase as determined by Interproscan

3.2.4.5. Comparison of supernatant proteome of *C. botulinum* A1 ATCC 19397 and *C. botulinum* B NCTC 7273

The extracellular proteomes of *C. botulinum* A1 ATCC 19397 and *C. botulinum* B NCTC 7273 were compared using local BLAST. This comparison revealed a core proteome of 100 homologous (BLAST E-value of $<1x10^{-30}$) proteins that were shared between the two strains, 18 proteins unique to NCTC 7273 and 98 proteins unique to ATCC 19397 (Figure 53).

The number of proteins belonging to 21 different functional groups that were detected in ATCC 19397 only, NCTC 7273 only and in both strains (core proteins) was determined (Table 29).

The size of the pan-proteome (i.e. the core and accessory proteome of ATCC 19397 and NCTC 7273) was established for each functional group of proteins. The similarity of the core proteome compared with the pan proteome was calculated for each functional group. For example, if every protein in a functional group was present in both ATCC 19397 and NCTC 7273 then there was 100% overlap between the pan and core proteome. The expression of proteins that are present in the core proteome is more highly conserved between the different strains, therefore these proteins are likely to be important in the ecological niche of the organism.

There were three functional groups where all identified proteins were present in both ATCC 19397 and NCTC 7273 i.e. a 100% overlap between the core and pan-proteome. These three groups were pathogenesis, adaptation to atypical conditions and signal transduction proteins. Pathogenesis had 6 proteins (BoNT, A7FS63; NTNH, A7FS62; HA33, A7FS60; HA17, A7FS59; HA70, A7FS58 and

clostripain, A7FUX5) while signal transduction contained two proteins (phosphotransferase systems) and adaptation to atypical conditions consisted of one protein (a cold shock protein).

Functional groups of proteins which had a core/pan-proteome overlap of more than 50% were protein fate (including thermolysin metalloprotease, A7FTX0 and chaperone protein DnaK, A7FXL5) , energy metabolism (including butyrate kinase, A7FZ18) nucleotide metabolism and transcription (RNA polymerase, A7FZ44).

Protein functional groups which had a core/pan-proteome overlap of 25-50% were amino acid biosynthesis, chemotaxis and motility (including flagellin, A7FWY1), detoxification (including desulfoferrodoxin, A7FYS9), sporulation and germination (Stage V sporulation protein S, A7FVX7), protein synthesis (including 30S ribosomal protein, AFZ55 and translation elongation factors, A7FPZ7), fatty acid and phospholipid metabolism and transport and binding proteins (ferritin family protein, A7FUJ4).

Protein functional groups with a low overlap between the core and pan-proteomes of 0-25% included proteins involved in biosynthesis of cofactors, cell envelope, central intermediary metabolism, DNA metabolism, mobile and extrachromosomal element functions, regulatory functions and hypothetical and unknown proteins (Table 29).

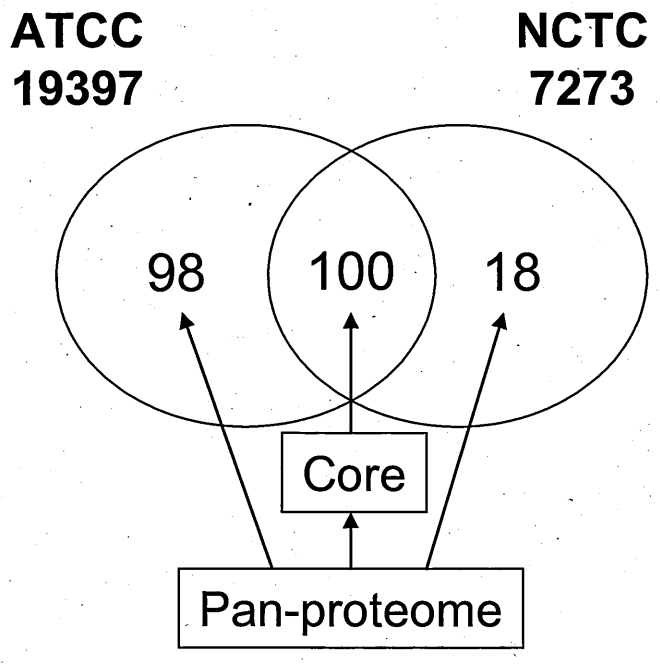


Figure 53: Core and unique proteins identified in 24 h culture supernatant of *C. botulinum* A1 ATCC 19397 and *C. botulinum* B NCTC 7273. The core and pan-proteomes are indicated.

Table 29: Comparison of the proteins identified in the supernatant of *C. botulinum* A1 ATCC 19397 and *C. botulinum* B NCTC 7273 by functional group. A protein is classified as belonging to the core proteome if it is present in the supernatant of both strains.

Protein functional group	Core	ATCC 19397 unique proteins	NCTC 7273 unique proteins	Pan proteome	Percentage core proteins
Adaptations to atypical conditions	1	0	0	1	100%
Amino acid biosynthesis	3	3	0	6	50%
Biosynthesis of cofactors, prosthetic groups, and carriers	1	4	0	5	20%
Cell envelope	2	7	1	10	20%
Central intermediary metabolism	1	4	1	6	17%
Chemotaxis and motility	3	2	1	6	50%
Detoxification	1	1	0	2	50%
DNA metabolism	0	1	0	1	0%
Energy metabolism	33	20	2	55	60%
Fatty acid and phospholipid metabolism	2	4	0	6	33%
Hypothetical and unknown proteins	10	32	6	48	21%
Mobile and extrachromosomal element functions	0	1	0	1	0%
Pathogenesis	6	0	0	6	100%
Protein fate	19	4	1	24	79%
Protein synthesis	7	8	2	17	41%
Purines, pyrimidines, nucleosides, and nucleotides	3	2	0	5	60%
Regulatory functions	0	1	1	2	0%
Signal transduction	2	0	0	2	100%
Sporulation and germination	1	0	1	2	50%
Transcription	3	2	0	5	60%
Transport and binding proteins	2	2	2	6	33%
Total	**100**	**98**	**18**	**216**	

3.2.5. Comparison of predicted and experimentally identified supernatant proteins of *C. botulinum* A1 ATCC 19397

Sub-cellular location of the protein products of the 3550 open reading frames in the *C. botulinum* A1 ATCC 19397 genome were predicted using five different tools (see section 3.1.3) (CELLO, LocateP, PsortB, SecretomeP and SignalP, for details see methods section 2.1.2). A consensus of the number of tools that predicted each protein to be extracellular was determined (summary in Table 30). This consensus was compared with the ATCC 19397 proteins that had been experimentally identified as extracellular at either 24 h or 96 h.

There were 11 proteins predicted by all five tools to be extracellular and four of these proteins (36.4%) were experimentally detected in the supernatant (see Table 30). Of the proteins predicted as extracellular by four tools 7.9% were experimentally identified as extracellular, of the proteins predicted by three tools 3.8% were extracellular, of the proteins predicted by two tools 4% were extracellular and of the proteins predicted by a single tool 9.7% were experimentally identified as extracellular. Of the 2994 proteins which were not predicted as extracellular by any tools, 5.5% were extracellular, a higher percentage than predicted by a consensus of two or three tools (Table 30).

Extracellular proteins predicted by each of the five tools were compared with the experimentally identified extracellular proteins; the percentage of predicted proteins that were detected in the supernatant was calculated. Prediction accuracy ranged between 4.9-15.7%, with the tool LocateP predicting extracellular proteins with the highest accuracy and SecretomeP predicting with the lowest accuracy (Table 31). The accuracy measurement was determined as the percentage of

experimentally observed proteins that were predicted to be extracellular by each tool.

Table 30: The number of *C. botulinum* A1 ATCC 19397 proteins predicted as extracellular by a consensus of 5, 4, 3, 2, 1 and 0 tools and the percentage of those proteins experimentally identified as supernatant proteins.

A consensus of this many tools predicted a protein to be extracellular	Number of proteins predicted to be extracellular	Number of experimentally identified extracellular proteins which were predicted to be extracellular	Percentage of predicted proteins that were experimentally identified as extracellular
5	11	4	36.4
4	113	9	7.9
3	235	9	3.8
2	124	5	4
1	72	7	9.7
0	2994	173	5.7

Table 31: The total number of predicted proteins, accurately predicted proteins and percentage of correct predictions by the 5 tools used to predict subcellular location when compared against the experimentally identified extracellular proteins.

Tool	Total number of predicted proteins	Total number of accurately predicted proteins	Specificity	Sensitivity
CELLO	460	27	5.90%	13.00%
LocateP	115	18	15.70%	8.70%
PsortB	156	10	6.40%	4.80%
SecretomeP	426	21	4.90%	10.10%
SignalP	379	24	6.30%	11.60%

3.2.6. Extracellular protein cost as an indicator of involvement in virulence

Extracellular proteins, once released by the cell, cannot usually be recovered and as such are a drain on the resources of the cell. Conversely, the constituent amino acids of intracellular proteins can be recycled upon protein degradation. It has been hypothesised that, because of their permanent loss to the cell, evolution has optimised extracellular proteins to reduce their synthetic burden on the cell (Smith & Chapman, 2010).

However, BoNT is a large protein that is released into the extracellular milieu, bucking the purported trend for extracellular proteins to be metabolically cheaper. It is also the single most important virulence factor possessed by *C. botulinum*. The hypothesis that extracellular proteins are metabolically more economical to produce than non-extracellular proteins was tested in *C. botulinum*. The relationship between cost of extracellular protein and involvement in virulence was also examined.

The average cost of synthesis of experimentally identified supernatant proteins was calculated (identified at 24 h and/or 96 h) and compared with the average synthesis cost of every coding sequence not detected extracellularly in the *C. botulinum* genome. The mean high-energy phosphate bond (HEPB) cost to synthesise an extracellular protein was 23.4 HEPBs per amino acid. The average cost for coding sequences not identified in the supernatant was 24.6 HEPBs per amino acid. The data sets were confirmed as normally distributed and a P-value of 3.00×10^{-25} obtained by t-test. This indicates that extracellular proteins are, on

average, cheaper than non-extracellular proteins by a small (4.9%) but significant margin.

BoNT, NTNH and the HA proteins were in the top 15% of either the total or average per amino acid cost. Three of the five toxin complex proteins were within the top three of the 207 proteins in terms of either total or average per amino acid cost (Table 32 & Table 33). Of the OrfX cluster proteins, OrfX1, OrfX2 and P-47 were in the top 15% of either the total or average per amino acid cost. OrfX3 was in the 25%.

Table 32: The *C. botulinum* A1 ATCC 19397 supernatant proteins with the highest total metabolic cost, BoNT toxin complex proteins are highlighted in red. Results presented include rank in regard to total cost, average HEPB cost per amino acid and rank in regard to average HEPB cost per amino acid. Cost is given in high energy phosphate bonds (HEPB).

Sequence ID	Sequence Name	Total Metabolic Cost (HEPB)	Sequence Length	Rank when extracellular proteins ordered by total HEPB cost	Average HEPB per AA	Rank when extracellular proteins ordered by average HEPB cost
A7FUJ2	Leucine rich repeat protein	32589.1	1359	1	24	56
A7FS63	BoNT/A	32143.2	1296	2	24.8	14
A7FS62	NTNH	29550.3	1193	3	24.8	15
A7FUC7	Collagenase	28974.4	1209	4	24	59
A7FZ77	DNA-directed RNA polymerase	28334	1232	5	23	142
A7FX10	Pyruvate-flavodoxin oxidoreductase	27734.8	1192	6	23.3	113
A7FZ76	DNA-directed RNA polymerase	27273.9	1178	7	23.2	123
A7FQC8	Isoleucine--tRNA ligase	26262	1038	8	25.3	5
A7FYQ6	Peptidase family protein	24097.1	975	9	24.7	18
A7FY60	N-acetylmuramoyl-L-alanine amidase	23416.8	967	10	24.2	37
A7FXV9	Penicillin-binding protein	21532.9	924	11	23.3	110
A7FVH5	LPXTG-motif cell wall anchor domain protein	20865.4	897	12	23.3	114
A7FQZ8	Aldehyde-alcohol dehydrogenase	20221.7	862	13	23.5	101
A7FR44	ClpB protein	20012.3	866	14	23.1	129
B1L2G0	OrfX2	17888.5	750	15	23.9	71
A7FWY3	Flagellar hook-associated protein	17866.6	811	15	22	186
A7FTG3	Putative xanthine dehydrogenase, molybdopterin-binding subunit	17713.1	765	16	23.2	122
A7FX72	Glycosyl hydrolase, family 18	17704.4	739	17	24	60
A7FV14	Carbohydrate binding protein	16192.6	682	18	23.7	79
A7FRB8	NlpC/P60 family protein	15982	718	19	22.3	177
A7FQA4	Methionyl-tRNA synthetase	15976	645	20	24.8	16
A7FZ72	Elongation factor G	15907	689	21	23.1	132
A7FY87	Threonine--tRNA ligase	15822.7	635	22	24.9	9
A7FWZ8	Chemotaxis protein CheA	15593.2	691	23	22.6	163
A7FW44	D-proline reductase, PrdA proprotein	15487.9	703	24	22	185
A7FW29	D-proline reductase, PrdA proprotein	15386.9	704	25	21.9	189
A7FV42	Chaperone protein htpG	15334.5	626	26	24.5	26
A7FS58	HA70	14924.8	626	27	23.8	69

Table 33: The *C. botulinum* A1 ATCC 19397 supernatant proteins with the highest average per amino acid metabolic cost, BoNT toxin complex proteins are highlighted in red. Results presented include rank in regard to total cost, average HEPB cost per amino acid and rank in regard to average HEPB cost per amino acid. Cost is given in high energy phosphate bonds.

Sequence ID	Sequence Name	Total Metabolic Cost (HEPB)	Sequence Length	Rank when extracellular proteins ordered by total HEPB cost	Average HEPB per AA	Rank when extracellular proteins ordered by average HEPB cost
A7FZF7	Acyl-ACP thioesterase	6636.6	249	126	26.7	1
A7FSQ9	Transcriptional regulator, MarR family	3982.3	156	172	25.5	2
A7FS59	HA17	3709.7	146	174	25.4	3
A7FUX2	Glutamate decarboxylase	11825.6	467	41	25.3	4
A7FQC8	Isoleucine--tRNA ligase	26262	1038	8	25.3	5
A7FQD5	Radical SAM domain protein	11637	460	43	25.3	6
A7FT23	Hypothetical protein	4198.9	166	166	25.3	7
A7FWQ1	Metallo-beta-lactamase family protein/flavodoxin	9779.5	389	66	25.1	8
A7FY87	Threonine--tRNA ligase	15822.7	635	22	24.9	9
A7FZ67	50S ribosomal protein L23	2416.2	97	190	24.9	10
A7FW59	Phosphopantetheine adenylyltransferase	4084.3	164	171	24.9	11
A7FQ93	Hypothetical protein	5046.1	203	149	24.9	12
B1L2G1	OrfX1	3524.9	142	184	24.8	14
A7FS27	thiamine pyridinylase	10022.2	404	58	24.8	13
A7FS63	BoNT	32143.2	1296	2	24.8	14
A7FS62	NTNH	29550.3	1193	3	24.8	15
A7FQA4	Methionyl-tRNA synthetase	15976	645	20	24.8	16
A7FSR0	Putative cyclase	4475.8	181	155	24.7	17
A7FYQ6	Peptidase family protein	24097.1	975	9	24.7	18
B1L2G3	P-47	10270.2		58	24.7	21
A7FYS9	Desulfoferrodoxin	3060.8	124	180	24.7	19
A7FXL0	HIT family protein	2808.8	114	184	24.6	20
A7FUJ4	Ferritin family protein	4198.1	171	167	24.6	21
A7FQ97	Polysaccharide deacetylase family protein	7094.4	289	118	24.5	22
A7FWP7	Sulfurtransferase	7944.9	324	105	24.5	23

A7FVH1	Metallopeptidase, family M24	14033.9	597	28	24.5	24
A7FZA5	Lysine--tRNA ligase	12353.2	504	37	24.5	25
A7FV42	Chaperone protein htpG	15334.5	626	26	24.5	26
A7FUS3	Hydrolase, NUDIX family	4356	178	158	24.5	27
A7FS60	HA33	7168.9	293	116	24.5	28

3.2.7. Identification of toxin complex proteins in clinical strains of *C. botulinum* by LC-MS/MS

The type and range of botulinum toxin complex proteins present in the supernatant of *C. botulinum* strains isolated from different clinical types of botulism was investigated by initial separation by SDS-PAGE followed by protein detection using LC-MS/MS. In order to maximise the number of strains that could be investigated, analysis was only undertaken on proteins extracted from targeted regions of the SDS-PAGE gel. To enable this approach, as protein products of the haemagglutinin and OrfX gene clusters are different in size, the toxin complex type of each clinical strain was determined by PCR prior to LC-MS/MS analysis (see section 3.3.2). To determine which areas of the gel should be targeted, initial experiments were carried out using supernatant protein from ATCC 19397 as an example of a haemaglutinin encoding strain and *C. botulinum* A3 Loch Maree (NCTC 2012) as an example of an OrfX encoding strain.

Approximately 40 protein bands were identified densitometrically on a 12% Bis-Tris SDS-PAGE gel. The gel was divided into 11 sections, taking into account both the total section size and the total protein density. These sections were analysed by LC-MS/MS and in total, 179 proteins were detected by more than one unique peptide. There was a maximum of 28 and a minimum of 9 proteins identified in these 11 gel fragments. Seventy of the 179 proteins were identified in the three sections between 55 kDa and 36.5 kDa.

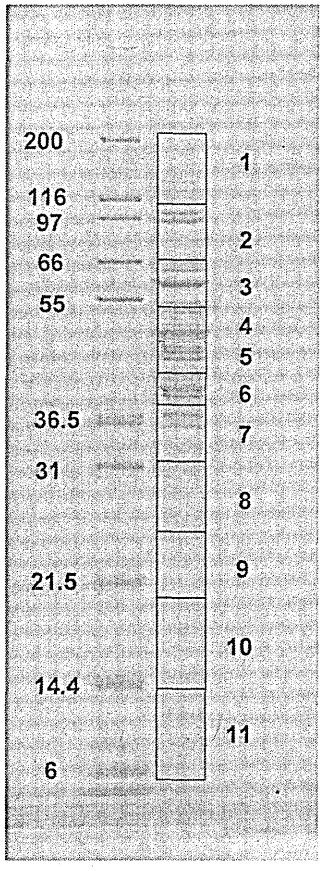
All the BoNT complex proteins were identified in ATCC 19397 by LC-MS/MS (Figure 54 A) and the unique peptides identified for each protein are shown in Table 26. The NTNH protein was identified in gel fragment 1 that contained proteins in the 200 kDa to 115 kDa range. This is consistent with the known 138

kDa molecular weight of NTNH. Although BoNT is 150 kDa, no peptides from this gel fragment matched BoNT. BoNT peptides were identified in the 115 kDa to 66 kDa section. When the BoNT peptides in this gel section were mapped to the toxin sequence they all matched the heavy chain of the toxin which has a molecular weight of 100 kDa. BoNT light chain (50 kDa) peptides were identified in the fourth gel fragment that included proteins of approximately 55 kDa to 50 kDa. Peptides matching HA70, were identified in the fifth gel fragment that included proteins of approximately 50 kDa to 42 kDa. All these peptides were localised to 60% of the protein on the C-terminus side. Peptides matching the HA33 protein were identified in the seventh gel fragment (36.5 kDa to 31 kDa). The ninth gel fragment (25 kDa to 20 kDa) contained peptides matching HA70. The peptides from this gel fragment mapped to 30% of the HA70 protein on the N-terminus side. HA17 peptides were identified in the tenth gel fragment that contained proteins from 20 kDa to 14.4 kDa in size. Based on these results, four SDS-PAGE gel fragments were analysed for each of the Haemagglutinin encoding clinical strains: between 150-95 kDa; 55-40 kDa; 37-30 kDa; 22-14 kDa.

Once the location on the SDS-PAGE gel of BoNT and the neurotoxin associated proteins from a haemagglutinin encoding *C. botulinum* strain was established, the same workflow was carried out for *C. botulinum* A3 Loch Maree (NCTC 2012) (Figure 54 B). The toxin gene for this strain is encoded within an OrfX positive toxin complex. All clinical strains were tested for their toxin complex type by PCR prior to LC-MS/MS analysis (for details see methods section 1.4.2 and results section 1.3.1) and the appropriate gel fragments for either the HA proteins or the OrfX proteins were analysed by LC-MS/MS depending on the results of the PCR. However, the use of *C. botulinum* A3 NCTC 2012 as reference for the production of the OrfX toxin complex proved to be inappropriate. Only NTNH and OrfX2 were

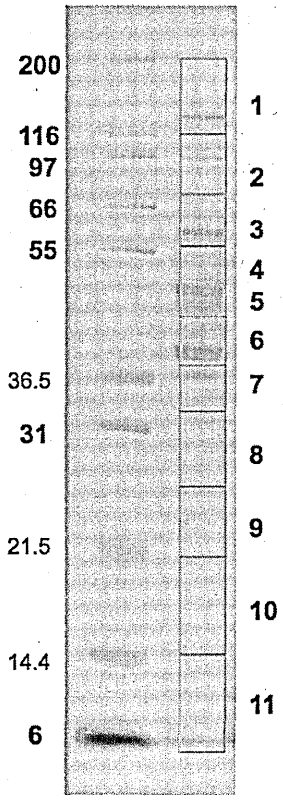
identified – both were in the appropriate gel fragments for their predicted molecular weight (Figure 54 B). Therefore, in addition to the NTNH and OrfX2 fragments, gel sections corresponding to the predicted molecular weight of BoNT (150 kDa), P-47 (47 kDa), OrfX1 (16.3 kDa) and OrfX3 (55.1 kDa) were analysed by LC-MS/MS. This resulted in three SDS-PAGE gel fragments being analysed for each of the OrfX positive clinical strains – the regions between 150-80 kDa, 60-40 kDa, and 22-14 kDa.

(A)



Gel frag	Proteins ided	BoNT	NTNH	HA70	HA33	HA17
1	9	N	Y	N	N	N
2	11	Y	N	N	N	N
3	13	N	N	N	N	N
4	22	Y	N	N	N	N
5	28	N	N	Y	N	N
6	20	N	N	N	N	N
7	17	N	N	N	Y	N
8	19	N	N	N	N	N
9	13	N	N	Y	N	N
10	10	N	N	N	N	Y
11	17	N	N	N	N	N

(B)



Gel frag	Proteins ided	BoNT	NTNH	P-47	OrfX1	OrfX2	OrfX3
1	8	N	Y	N	N	N	N
2	10	N	N	N	N	Y	N
3	9	N	N	N	N	N	N
4	19	N	N	N	N	N	N
5	14	N	N	N	N	N	N
6	17	N	N	N	N	N	N
7	22	N	N	N	N	N	N
8	17	N	N	N	N	N	N
9	21	N	N	N	N	N	N
10	18	N	N	N	N	N	N
11	23	N	N	N	N	N	N

Figure 54: Identification of the SDS-PAGE gel fragments that contain the botulinum toxin and toxin complex proteins by LC-MS/MS for (A) *C. botulinum* A1 ATCC 19397 24 h supernatant and (B) *C. botulinum* A3 NCTC 2012 24 h supernatant

Table 34: The unique peptides identified for each botulinum toxin complex protein in ATCC 19397 supernatant protein at 36 h

Protein	Number of unique peptides	Unique peptide sequences
NTNH	13	EFSIMMPDR FLQAIITLLK FSLSSDFVEVVSSK LDEVIISVLDNMEK LNSLISSTIPFPYAGYR MNINDNLSINSPVDNK NIYETEIEGNNAIGNDIK SLFSSETALLIK SVLAQETLIK VAPNIWVAPER VDGGIYDSNFLSQDSEK YDEFYIDPAIELIK YYGESLSIDEEYK
BoNT Hc	11	AIINYQYNQYTEEEK IALTNSVNEALLNPSR LLSTFTEYIK NIINTSILNLR NNINFNIDDLSSK NQIQLFNLESSK VLTVQTIDNALSK VNNTLSTDIPFQLSK WIFVTITNNR YFNLFDK YSQMINISDYINR
BoNT Lc	6	FIDSLQENEFR MLTEIYTEDNFVK NFTGLFEFYK QVPVSYYDSTYLSTDNEK SIVGTTASLQYMK VIDTNCINVIQPDGSYR
HA70 - 50 kDa	11	AINYITGFDSPNAK DAFNVQLFNTSTSLFK INAQNNLPSLK IYEAIGSGNR LLNGAIYILK LYTSYNQGIGTLFR SYLVVLLNK TNDKDLIGTLLIEAGSSGSIIQPR VEVTELNNYNIR VPQTSSNIENQIQFK YELIDYQNGSIVNK
HA70 - 20 kDa	7	AVLYVPSLGYVK CILNEQFLYK DFYFLTNDK NLYMYLQYTYIR QNQILGGSVISNGSTGIVGDLR SIEFNPGEK VINYSDTIDLADGNYVVR
HA33	4	DIGNNSFIIASYK ISPILDLNK LSTLNNSNYIK NPNLVLYADTVAR
HA17	3	SIFSGSLYLNPVSK TFLPNGNYNIK WNVEYMAENR

Table 35: The unique peptides identified for each botulinum toxin complex protein in NCTC 2012 supernatant protein at 36 h

Protein	Number of unique peptides	Unique peptide sequences
NTNH	5	GTMDNFYAAYK
		INNNFNIDSPVDNK
		NSDPFIPVYNITETK
		SVLAQESLIK
		YYGESLNINEDQK
OrfX2	8	CIIDDGYLDMNFGTSSEK
		ENYSINEIIPK
		IAVEDAGLISDDGTTSIR
		LQGIYLLGGALEK
		NDSGVTDIELQEINR
		NVDSFLKPGK
		QEYELALESK
		SWNLTDEGEGSHPVLK

Once the gel fragments that contained BoNT and its associated proteins were identified the corresponding gel fragments from clinical isolates of *C. botulinum* were analysed.

Supernatant proteins were precipitated after 36 h growth in liquid culture from 22 clinical strains of *C. botulinum* and analysed for the presence of BoNT and associated proteins using 1DGE-LC-MS/MS (as described in methods section 1.3.2-1.3.4). These strains included 15 wound botulism isolates, five infant botulism strains and two food botulism strains. There were 14 type A strains, six type B strains and two bivalent AB encoding strains. The National Botulinum Reference Laboratory at the Health Protection Agency had previously determined and documented toxin serotype for these isolates using qPCR and mouse neutralisation bioassay (Akbulut et al., 2004). Toxin complex type was determined prior to LC-MS/MS according to methods detailed in Materials and Methods section 1.4.2. The LC-MS/MS analysis was targeted to the gel fragments where the toxin complex proteins had been identified previously (as described above).

The targeted LC-MS/MS approach was used to detect and identify the BoNT complex proteins in 22 clinical isolates – the correct BoNT serotype was detected and identified in 20/22 strains (Table 36). At least 2 toxin complex proteins were identified in every strain analysed with all the HA toxin complex proteins (neurotoxin, NTNH, HA70, HA33 and HA17) being identified in 13/22 strains.

There were 15 wound botulism strains analysed: 10 type A strains, 4 type B and 1 type AB. All 15 strains encoded genes for the HA toxin complex type. BoNT, NTNH and all the HA toxin complex proteins (HA33, HA17 and HA70) were identified in 10 of these 15 strains. There were different patterns of protein

absence in the five strains that lacked at least one member of the toxin complex (Table 36). BoNT, NTNH and HA70 were not identified in the supernatant of 04068341, a type A wound botulism strain. BoNT and NTNH were not identified in the supernatant of H091280045, a type B wound botulism strain. HA70 was not detected in the supernatant of H065060505, a type A wound botulism strain. HA17 was not identified in the supernatant of H091140481 or H102120680, type B strains which caused wound botulism.

Five infant botulism strains were analysed: three type A strains (H040660361, H094460264 and H112480657) and two type B strains (H074400585 and H090840606). HA toxin complex genes were detected in three of these isolates (two type B and one type A), OrfX genes were identified in one strain (type A) and both HA and OrfX genes were identified in one strain (type A) (Table 36). LC-MS/MS detected both BoNT and NTNH in the supernatant of all these isolates. All three HA toxin complex proteins were identified in the three HA positive infant botulism isolates. Only OrfX2 was identified in the supernatant of the OrfX positive isolate. In the supernatant of the dual HA and OrfX positive infant botulism isolate, OrfX2 and OrfX3 were detected but no HA proteins.

Two strains isolated from clinical food botulism cases were investigated: one type A, HA encoding isolate (H114580650) and one bivalent AB isolate, that encoded both HA and OrfX genes (H063740588). BoNT/A and NTNH were identified in the supernatant of both these strains, BoNT/B was not identified in the supernatant of the bivalent strain. In the supernatant of the HA positive isolate, HA70 and HA33 were detected. In the supernatant of the bivalent strain only HA70 and HA33 were detected.

The type strain *C. botulinum* A1(B) NCTC 2916 was also examined as part of this investigation. In this strain the type A toxin is present in an OrfX toxin complex cluster and the type B toxin is present in an HA cluster. Type A toxin protein was detected in the culture supernatant of this isolate, while only the HA toxin complex proteins (HA70, HA33 and HA17) were detected.

The presence of clostripain, an enzyme putatively involved in nicking the toxin resulting in the formation of the active dichain was also investigated. It was identified in the supernatant of 20/22 strains, the two strains for which clostripain was not identified were associated with foodborne botulism. Both of these organisms showed the presence of BoNT in the 50 kDa region of the gel indicating that despite the absence of clostripain BoNT was still being cleaved into a dichain.

Table 36: Detection and identification of toxin associated proteins in the supernatant of *C. botulinum* clinical isolates. Toxin type was determined by qPCR and toxin complex type was determined by PCR.

Lab ID	Type of botulism	Toxin type	Toxin complex	Bont A	BoNT/B	NTNH	HA70	HA33	HA17	P-47	OrfX1	OrfX2	OrfX3	Clostripain
ATCC 19397	Type strain	A1	HA	√	X	√	√	√	√	X	X	X	X	√
NCTC 2012	Type strain - FB	A3	OrfX	X	X	√	X	X	X	X	X	√	X	√
NCTC 2916	Type strain	A1(B)	HA/OrfX	√	X	√	√	√	√	X	X	X	X	√
H112480657	IB	A	OrfX	√	X	√	X	X	X	X	X	√	X	√
H094460264	IB	A	HA/**OrfX**	√	X	√	X	X	X	X	X	√	√	√
H063740588	FB	AB	**HA**/OrfX	√	X	√	√	√	X	X	X	X	X	X
H114580650	FB	A	HA	√	X	√	√	√	X	X	X	X	X	X
H040660361	IB	A	HA	√	X	√	√	√	√	X	X	X	X	√
H074400585	IB	B	HA	X	√	√	√	√	√	X	X	X	X	√
H090840606	IB	B	HA	X	√	√	√	√	√	X	X	X	X	√
H040680341	WB	A5(B)	HA	X	X	X	X	√	√	X	X	X	X	√
H042440055	WB	A5(B)	HA	√	X	√	√	√	√	X	X	X	X	√
H044020065	WB	A5(B)	HA	√	X	√	√	√	√	X	X	X	X	√
H044640107	WB	A5(B)	HA	√	X	√	√	√	√	X	X	X	X	√
H052880114	WB	A	HA	√	X	√	√	√	√	X	X	X	X	√
H063960325	WB	A	HA	√	X	√	√	√	√	X	X	X	X	√
H064620409	WB	A	HA	√	X	√	√	√	√	X	X	X	X	√
H065060505	WB	A	HA	√	X	√	X	√	√	X	X	X	X	√
H071040476	WB	A	HA	√	X	√	√	√	√	X	X	X	X	√
H075000578	WB	A	HA	√	X	√	√	√	√	X	X	X	X	√
H091640054	WB	AB	HA	√	X	√	√	√	√	X	X	X	X	√
H062260493	WB	B	HA	X	√	√	√	√	√	X	X	X	X	√
H091140481	WB	B	HA	X	√	√	√	√	X	X	X	X	X	√
H091280045	WB	B	HA	X	X	X	√	√	√	X	X	X	X	√
H102120680	WB	B	HA	X	√	√	√	√	X	X	X	X	X	√

3.2.8. Summary of findings of proteomic investigation

- Endopeptidase assay reported decline in active supernatant toxin between 24 h and 96 h, in contrast to previous reports.
- The *C. botulinum* supernatant proteome is more complex than previously reported, with a larger number of proteins identified.
- The *C. botulinum* supernatant proteome is stable between 24 h and 96 h, although there are signs of protein degradation.
- Significantly more proteins were identified in the supernatant proteome of BoNT/A producing ATCC 19397 than BoNT/B producing NCTC 7273.
- All sub-cellular localisation prediction tools have low sensitivities (best was 15.7%) and specificities (best was 13%) at predicting *C. botulinum* supernatant proteins.
- Novel, potential virulence associated proteins were identified in the *C. botulinum* supernatant.
- *C. botulinum* extracellular proteins involved in virulence are some of the metabolically most expensive proteins identified in the supernatant.
- All 22 clinical strains of *C. botulinum* produced at least one component of the ANTPs, while the majority of strains produced all ANTPs, BoNT and NTNH.

3.3. Genomic diversity and toxin complex type of clinical isolates causing botulism in the UK

Botulinum neurotoxin genes are encoded alongside and co-expressed with genes encoding the associated non-toxic proteins. There are two known toxin complex types, the haemagglutinin (HA) and the OrfX. The HA complex type has been more thoroughly characterised; isolates from the majority of cases of botulism have been found to encode the HA complex type. Although most strains of *C. botulinum* express a single toxin type, some strains produce more than one type. Usually, one type of toxin predominates and this is usually denoted by use of a capital letter with the toxin type produced to a lesser extent denoted in lower case. Strains that encode two toxin types but only produce one type are denoted A(B), where type A is produced and B is not. Some bivalent strains are also bivalent for toxin complex type (i.e. encode genes for both HA and Orfx) with the toxin genes being associated with different toxin complex types. Toxin types A1, A5, B, C and D have been found in *ha* clusters while A1 (rarely), A2, A3, A4, Bf, E and F type toxins have been found in *orfX* clusters.

The toxin complex type of isolates from cases of botulism in the UK is not routinely determined by the National Reference Laboratory. In this study, the relationship between toxin complex type, genomic similarity and type of botulism is investigated. This involved developing a PCR-based assay to determine toxin complex type and applying fluorescent Amplified Fragment Length Polymorphism (fAFLP) to provide insight into the genetic diversity of isolates causing botulism in the UK. Seven strains from the National Collection of Type Cultures, representing a range of toxin and toxin complex types were employed as reference strains.

3.3.1. Investigating genomic diversity of isolates of causing different forms of botulism in the UK

fAFLP was employed to investigate the genomic diversity of 38 isolates of proteolytic *C. botulinum* and neurotoxigenic *C. butyricum* from clinical cases of botulism collected between 2004-2011 and from the National Collection of Type Cultures. Three technical replicates of each isolate were processed to ensure reproducibility. The average level of similarity of technical replicates was 94.9%, with a range of 91.4-96.9%. Isolates with fAFLP profile similarity greater than 90% were treated as indistinguishable.

Fragments ranging in size between 60-600 bp were analysed for each strain and a dendrogram representing the diversity of fAFLP profiles was generated. The strains grouped into five main clusters (Figure 55) that shared more than 60% fAFLP profile similarity. With the exception of four strains, all strains examined grouped into one of these five clusters. There were also sub-clusters within the 5 main clusters that showed at least 80% similarity to each other. All four of the non-clustered isolates were group I *C. botulinum* and consisted of 3 type strains isolated in 1947 (*C. botulinum* B NCTC 7273), 1976 (*C. botulinum* Ba4 657) and 1922 (*C. botulinum* A3 NCTC 2012) and a BoNT/B producing organism (H074400586) which was linked to a case of infant botulism in 2007.

All *C. butyricum* strains analysed here formed a distinct cluster (Cluster 5) with only 22% similarity to the group I *C. botulinum* strains (clusters 1-4). Within Cluster 5 there was a sub-cluster of four indistinguishable strains (fAFLP profile >90% similar), three of these strains were associated with a single case of infant botulism, representing an isolate from the afflicted infant, the infant's mother

(asymptomatic carriage) and an environmental isolate (from the family's terrapin tank). The fourth strain was another toxigenic *C. butyricum* strain isolated from an unrelated infant botulism case. The fifth strain in cluster 5, NCTC 7423, is a non-toxigenic *C. butyricum* type strain, which was significantly different (fAFLP profile <90% similar) from the other Cluster 5 strains.

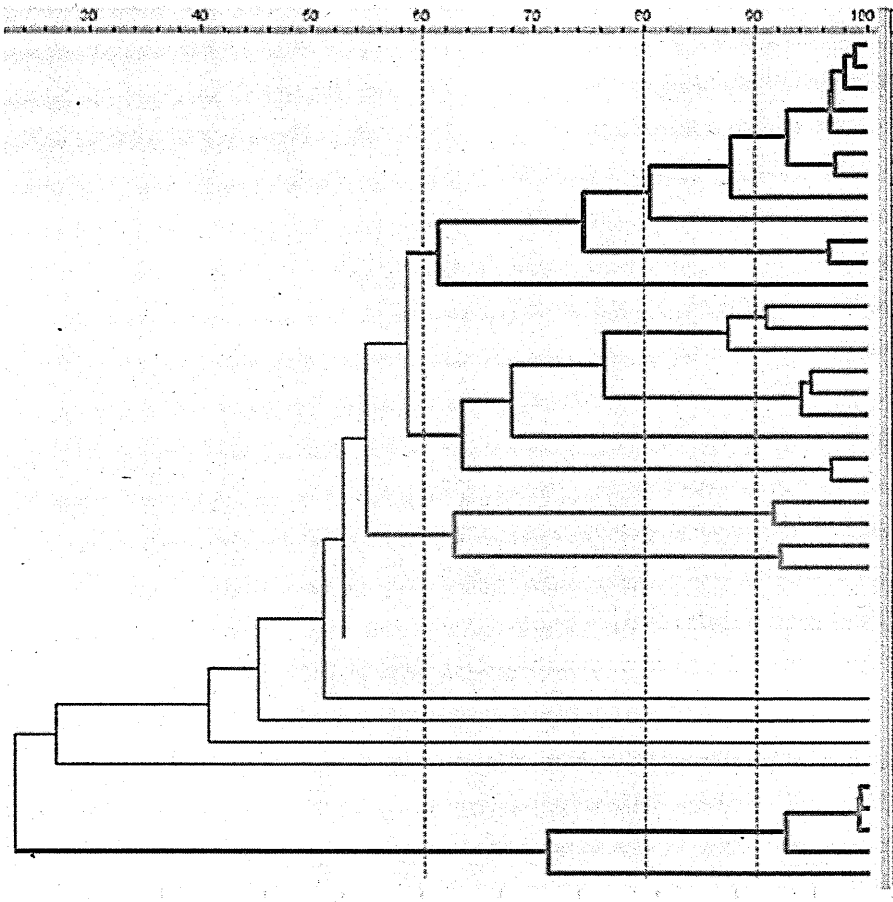
The *C. botulinum* group I isolates analysed clustered into four different groups that had more than 60% similarity with the exception of the four strains detailed above (three type strains and H074400586) (Figure 55). Cluster 1 contained 12 strains; 9 type A strains and 3 type B strains. These strains were associated with 11 cases of wound botulism (9 type A strains, 2 type B strains) and 1 case of infant botulism (a type B strain). There were two sub-clusters (>80% similarity) in Cluster 1, one of which contained all nine of the type A encoding strains in Cluster 1. The other sub-cluster contained two of the three type B strains in Cluster 1. The only infant botulism strain in Cluster 1 was a type B encoding strain that was also the only strain not to fall within a sub-cluster.

Cluster 2 consisted of 9 strains; 3 type A, 4 type B and 2 bivalent AB strains. There were 5 wound botulism strains (2 type A, 3 type B), 1 infant botulism strain (type B) and 2 bivalent AB food borne botulism strains. Three sub-clusters within Cluster 2 contained more than one isolate. In addition to grouping by fAFLP profile, these sub-clusters were also consistent in their respective toxin types. One sub-cluster consisted of three type A strains, another consisted of three indistinguishable type B strains while the final sub-cluster contained two indistinguishable type AB strains. One of the bivalent AB strains was a type strain isolated from canned corn in the USA in 1929, while the other was associated with

food botulism in 2006. One Cluster 2 strain did not share 80% similarity with any other isolate this was a type B strain associated with infant botulism in 2011.

Cluster 3 consisted of 4 strains, all type A, that grouped into two pairs of indistinguishable (>90% similarity) strains. The first of these pairs consisted of the type A encoding National Collection of Type Culture strains NCTC 7272 and ATCC 19397, these strains were derived from the same initial culture. The other pair of isolates were from a patient and environmental sample from the same food-borne outbreak, associated with commercially prepared korma sauce in 2011.

Cluster 4 consisted of 5 strains from different cases, all 5 of which encoded type A toxin, all associated with cases of infant botulism between 2007 and 2011. There were two sub-clusters with more than one isolate within Cluster 5, with a single strain not showing 80% similarity to any other strain.



Key	Toxin type	Type of botulism	Toxin complex type	Year	Location
H040680341	A5(B)	WB	HA	2004	London
H052880114	A	WB	HA	2005	Dorset
H071040476	A	WB	HA	2007	Bournemouth
H044020065	A5(B)	WB	HA	2004	Dudley
H103660204	A	WB	HA	2010	Birmingham
H042440055	A5(B)	WB	HA	2004	Sheffield
H065060505	A	WB	HA	2006	Winchester
H075000578	A	WB	HA	2007	Bristol
H044640107	A5(B)	WB	HA	2004	Ashton-under-lyne
H091140481	B	WB	HA	2009	Halifax
H091280045	B	WB	HA	2009	Kettering
H090840606	B	IB	HA	2009	Greece
H040660361	A	WB	HA	2004	Cardiff
H063960325	A	WB	HA	2006	Liverpool
H071400014	A	WB	HA	2007	Gloucester
H092080005	B	WB	HA	2009	York
H093620104	B	WB	HA	2009	Cambridge
H062260493	B	WB	HA	2006	Kirkcaldy
H114400598	B	IB	HA	2011	London-Noah
NCTC 2916	A1B	Type-env	HA/OrfX	1928	
H063740588	AB	FB	HA/OrfX	2006	Mayo, Ireland
NCTC 7272	A	Type	HA	1934	
ATCC 19397	A	Type	HA	1934	
H114590007	A	FB-env	HA	2011	Glasgow
H114680633	A	FB	HA	2011	Glasgow
H093320637	A	IB	OrfX	2009	London
H111860974	A	IB	OrfX	2011	Romford
H112480657	A	IB	OrfX	2011	London
H074240406	A	IB	OrfX	2007	London
H094460264	A	IB	HA/OrfX	2009	Edinburgh
NCTC 7273	B	Type	HA	1947	
H074400586	B	IB	HA	2007	London
Ba4 657	Ba4	Type	HA/OrfX	1976	
NCTC 2012	A	Type-FB	OrfX	1922	
H110141052	Butyricum-E	IB		2011	Dublin-Oliv
H110220838	Butyricum-E	IB		2011	Dublin-Kris
H110340631	Butyricum-E	IB-env		2011	Dublin
H102020561	Butyricum-E	IB		2010	Stoke-on-trent
NCTC 7423	Butyricum-E	Type		1948	

Figure 55: Proteolytic *C. botulinum* and *C. butyricum* clustering based on fAFLP analysis with information on toxin type, type of botulism (WB = wound botulism, IB = infant botulism, FB = food botulism) and toxin complex type (determined in section 3.3.2). Strains highlighted in green = cluster 1, red = cluster 2, dark blue = cluster 3, yellow = cluster 4 and light blue = cluster 5. There are 3 dashed lines, at 60% (cluster), 80% (sub-cluster) and 90% (indistinguishable) similarity.

3.3.2. Determination of the toxin complex type of clinical strains of C. botulinum

The toxin complex type (*ha* or *orfX*) of the 32 group I clinical strains associated with separate incidents of botulism were determined using PCR assays to detect *ha70*, *ha33*, *orfX/2* and *orfX/3* (Table 37). Only group I strains were investigated as these cause the majority of botulism in the UK and to analyse toxigenic *C. butyricum* strains here would require the use of different primer sets due to the variation between their *orfX* sequences (toxin type E has always been found in *orfX* toxin complex types) and the *orfX* sequences of group I strains.

Of the 34 clinical *C. botulinum* strains analysed, 28 were positive for only *ha* encoding genes, 4 were positive for only *orfX* encoding genes and 2 were positive for both *ha* and *orfX* genes. Among the 28 *ha* only strains, 24 were positive for both *ha70* and *ha33* with two testing positive for *ha33* but negative for *ha70*. Among the 4 clinical *orfX* only encoding strains, 3 were positive for both *orfX/2* and *orfX/3* while 1 was negative for *orfX/2* and positive for *orfX/3*. Both the clinical *ha* and *orfX* positive strains tested positive for *ha70, ha33, orfX/2* and *orfX/3*.

Of the four food botulism isolates from different incidents, three were *ha* only and one was bivalent *ha/orfX*. Of the eight infant botulism isolates, three were *ha* only, one was bivalent *ha/orfX* and four were *orfX* only. All 21 wound botulism strains were *ha* only encoding strains.

Two strains exhibited unusual combinations of genes. H091640054 was bivalent in terms of toxin type, encoding both BoNT/A and B, however it was monovalent in terms of toxin complex type, encoding only HA. On the other hand H094460264

encoded only BoNT/A but was positive by PCR for genes from both the HA and OrfX complex types.

Table 37: Clinical strains which were analysed for toxin complex type and toxin serotype, type of botulism caused and toxin complex type as determined by PCR. Strains marked with + came from the same incident.

Molis Number	HA70	HA33	OrfX2	OrfX3	Toxin complex type	Type of botulism	Toxin type
H040660361	+	+	-	-	HA	WB	A
H040680341	+	+	-	-	HA	WB	A5(B)
H042440055	+	+	-	-	HA	WB	A5(B)
H044020065	+	+	-	-	HA	WB	A5(B)
H044640107	+	+	-	-	HA	WB	A5(B)
H052880114	+	+	-	-	HA	WB	A
H062260493	+	+	-	-	HA	WB	B
H063740588	+	+	+	+	HA/OrfX	FB	AB
H063960325	+	+	-	-	HA	WB	A
H064620409	+	+	-	-	HA	WB	A
H065060505	-	+	-	-	HA	WB	A
H065260139	+	+	-	-	HA	WB	A
H071040476	+	+	-	-	HA	WB	A
H071400014	+	+	-	-	HA	WB	A
H074240407	-	-	+	+	Orf	IB	A
H074400585	+	+	-	-	HA	IB	B
H075000578	+	+	-	-	HA	WB	A
H090840606	-	+	-	-	HA	IB	B
H091140481	+	+	-	-	HA	WB	B
H091280045	+	+	-	-	HA	WB	B
H091640054	+	+	-	-	HA	WB	AB
H092080005	+	+	-	-	HA	WB	B
H093320637	-	-	-	+	OrfX	IB	A
H093620104	+	+	-	-	HA	WB	B
H094460264	+	+	+	+	HA/OrfX	IB	A
H111860974	-	-	+	+	ORF	IB	A
H112120680	+	+	-	-	HA	WB	B
H112480657	-	-	+	+	ORF	IB	A
H113660204	+	+	-	-	HA	WB	A
H114400598	+	+	-	-	HA	IB	B
H114580650	+	+	-	-	HA	FB	A
H114580654	+	+	-	-	HA	FB	A
H114590007 +	+	+	-	-	HA	FB - env	A
H114680633 +	+	+	-	-	HA	FB	A
ATCC 19397	+	+	-	-	HA	Type strain	A1
NCTC 2916	+	+	+	+	HA/OrfX	Type strain	A1(B)
NCTC 2012	-	-	+	+	OrfX	Type strain - FB	A3

3.3.3. Summary of findings of genomic diversity, toxin complex type and form of botulism

- There is considerable diversity among the clinical and type strains investigated in the fAFLP experiment.
- There is a relationship between toxin complex type, genomic background and type of botulism. Strains that encode OrfX and HA fall into different fAFLP clusters. These strains are also associated with different forms of disease, with all the OrfX encoding strains being associated with infant botulism.
- There are two strains with novel combinations of toxin and toxin complex type. H091640054 encodes both BoNT/A and B while only encoding the HA toxin complex type. H094460264 encodes both HA and OrfX while encoding only BoNT/A.

3.4. RNA-Seq transcriptome profiling to investigate C. botulinum toxin complex expression

RNA-seq was used to investigate transcription of the genes encoding the botulinum toxin complex. Initial experiments with *C. botulinum* A1 ATCC 19397 were undertaken using reverse transcription quantitative PCR (RT-qPCR) to determine the expression profile of *bont/A*, over a time course, relative to a reference housekeeping gene - *gluD* (encoding glutamate dehydrogenase). Three samples from pre-peak, peak and post-peak toxin relative gene expression were analysed by RNA-seq. Insight into *bont/A* regulation and details of transcription from both coding and non-coding regions of the genome are reported.

3.4.1. Optimisation of cell lysis and RNA extraction

Preliminary experiments to determine a methodology for consistently providing sufficient quantity of high quality RNA for downstream analysis were performed.

Optimisation of lysis and extraction was carried out using *C. botulinum* A1 ATCC 19397 cultures grown anaerobically in TPGY at 37°C to mid-log phase. Two methods of cell lysis and two methods of RNA-extraction (giving four different combinations of lysis and extraction methods) were assessed. For cell lysis, bead beating and enzyme lysis were investigated. The RNA extraction methods used were phenol:chloroform:isoamyl alcohol treatment and silica column extraction (RNeasy, Qiagen). The quantity of RNA was measured using the Quant-iT Ribogreen kit (Life Technologies, Paisley, UK) while RNA quality was determined by Bioanalyser (Agilent, Wokingham, UK) (Table 38 & Figure 56).

The highest quantity of RNA (373.5 ng/µl) was obtained using bead beating lysis and extraction phenol:chloroform:isoamyl alcohol, however this RNA was of a low quality, showing significant degradation of the 16S and 23S rRNA (Figure 56 (A)). The second highest yield (170.1 ng/µl) was obtained by lysing cells with enzyme treatment and extracting RNA by a phenol:chloroform:isoamyl alcohol method. However, this RNA also showed signs of degradation. Bead beating lysis followed by column extraction gave a very low yield (3.4 ng/µl) of low quality RNA. Enzyme lysis followed by column extraction gave an acceptable yield of 104.3 ng/µl and an excellent RNA quality (Figure 56). Therefore the combination of enzyme lysis and column extraction was employed for further RNA extraction.

The RNeasy columns used for extraction do not capture RNA molecules less than 200 nucleotides in length. In order to capture these RNAs an additional extraction step was carried out in which flow through from the RNeasy columns was collected and a phenol:chloroform:isoamyl alcohol RNA extraction performed. This RNA was then combined with the RNA captured by the column to provide a high quality extract of total cellular RNA (Figure 57).

Table 38: Quantity of RNA extracted from an 8 h culture of *C. botulinum* A1 19397 using different combinations of lysis and RNA extraction methods.

		RNA extraction method	
		Column	Phenol:chloroform:isoamyl alcohol
Lysis method	Bead beating	3.4 µg/µl	373.5 µg/µl
	Enzyme	104.3 µg/µl	170.1 µg/µl

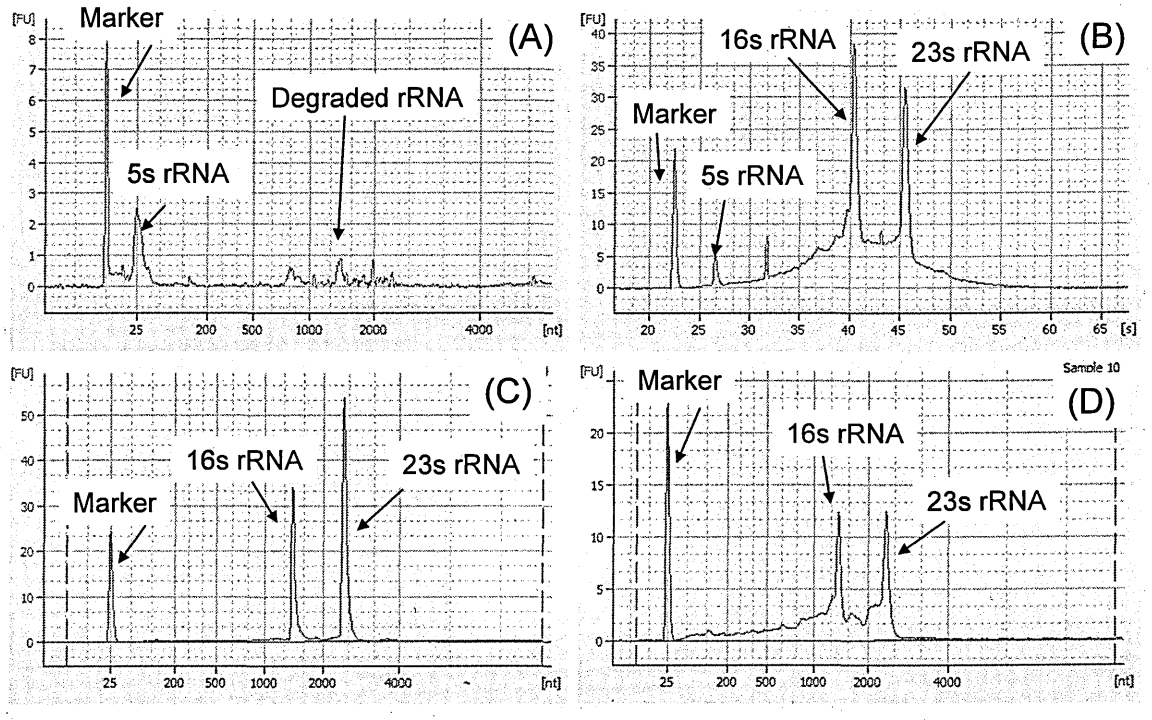


Figure 56: Assessment of RNA quality by Bioanalyser, high concentration of intact rRNA indicates good quality. RNA quality was assessed following extraction by (A) bead beating followed by column extraction (B) bead beating and phenol:chloroform extraction (C) Enzyme lysis followed by column extraction (D) Enzyme lysis followed by phenol:chloroform extraction. [FU] = arbitrary fluorescence units [nt] = fragment size, nucleotides.

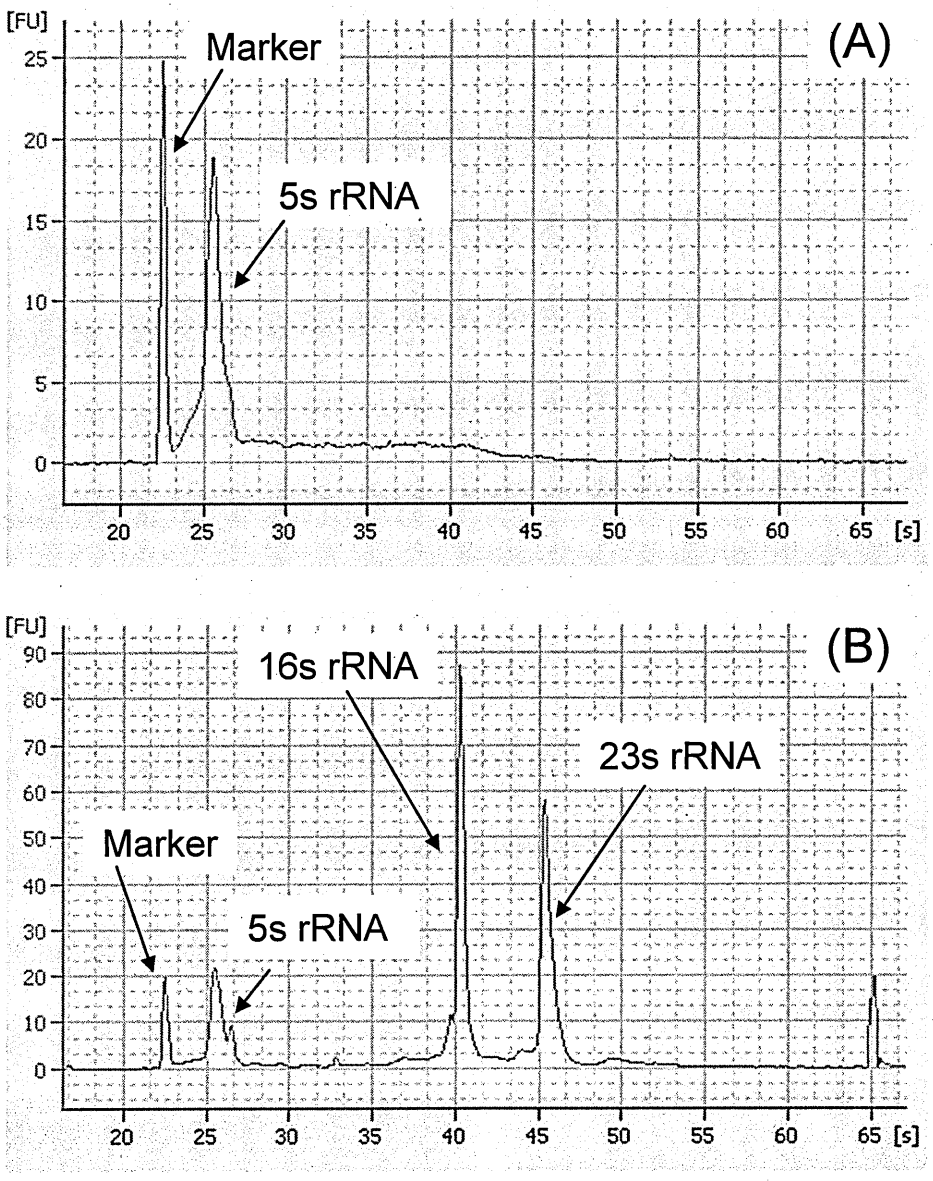


Figure 57: Total cellular RNA was obtained by combining RNeasy column captured RNA with small RNAs captured from the column flow through by phenol:chloroform extraction (A) RNA extracted from the flow through only(B) Total RNA from flow through and column.

3.4.2. RT-qPCR investigation of *bont/A* gene expression relative to the reference gene *gluD* along a time course

Previous work has shown that *bont/A* expression peaks at late log or early stationary phase (Shin et al., 2006). Therefore, a growth curve was carried out to determine the time point at which broth cultures of *C. botulinum* ATCC A1 19397 reached log and early stationary growth phases. The OD_{600} of an ATCC 19397 culture was measured at 0, 4, 6, 7, 8, 9, 10, 11, 12 and 13 h. The cultures entered logarithmic growth between 4-6 h, undergoing rapid growth until 10-12 h when the OD_{600} plateaued as the cultures entered stationary phase. The results obtained closely mirror those obtained during the proteomic workflow (see section 3.2.1).

The expression of *bont/A* relative to the expression of the reference gene *gluD* (encodes glutamate dehydrogenase, Gene ID 5395388) was determined by quantitative PCR (qPCR) analysis of reverse transcribed RNA (cDNA) taken at hourly intervals from 7 h to 13 h and at 24 h from an ATCC 19397 broth culture. Before the gene expression of *bont/A* relative to *gluD* could be calculated, the efficiency of the two reverse transcription qPCRs (RT-qPCR) was determined. Five ten-fold dilutions of total RNA were tested in triplicate using the RT-qPCR assays for *bont/A* and *gluD* and the threshold cycle (C_t) for each dilution determined. The C_t and log-RNA concentration were plotted on a graph and the slope of the line (Figure 58) used to derive the efficiency of the reaction according to (see 2.5.8). The efficiency of the *bont/A* RT-qPCR was 111.2% and the efficiency of the *gluD* RT-qPCR was 90.9%.

For each cDNA sample the *bont/A* qPCR and the *gluD* qPCR C_t values were determined. Relative gene expression of *bont/A* compared with *gluD* was then

calculated (Figure 59) for each time point using the method of Pfaffel, 2006, which takes the efficiency of the RT-qPCR into account.

The relative expression of *bont/A* compared to *gluD* at 7 h is given a base line value of 1 and expression at later time-points is normalised to this baseline. The relative expression of *bont/A* in sample A increased 7.6 fold between 7 h and 9 h, when it peaked, before gradually decreasing until 24 h when relative expression was less than 10% of the 7 h level. In sample B an increase of 6 fold between 7 h and 10 h was followed by a rapid decrease of 3.2 fold between 10 h and 11 h, stabilising between 11 h and 13 h before decreasing to less than 10% of 7 h expression at 24 h.

The reference gene *gluD* was chosen because of concerns about the most commonly used reference gene for relative gene expression experiments in *C. botulinum*, 16S rRNA. These concerns included potential changes in the rate of synthesis of rRNA as cell growth slows (Hansen et al., 2001) and the fact that it has been proposed to interfere with the calculation of the relative gene expression of *bont* (Lovenklev et al., 2004). Specifically, *gluD* was chosen because it had been previously used in relative gene expression experiments of *Clostridium difficile* (Kirk et al., 2014).

Three time-points from each sample, representing the *C. botulinum* A1 ATCC 19397 transcriptome before peak toxin expression, at peak toxin expression and after peak toxin expression were selected for analysis using RNA-seq. For sample A these were the 7 h, 9 h and 13 h time-points and for sample B the 7 h, 10 h and 13 h time-points were selected. These samples are hereafter referred to as the mid-log (7 h), late log (9 h & 10 h) and early stationary phase (13 h) samples. The

amount of 5S, 16S and 23S rRNA in the samples was reduced by Ribo-Zero rRNA reduction kit (CamBio, Cambridge, UK) (Figure 60) and the samples were then sent for RNA-seq at an external facility.

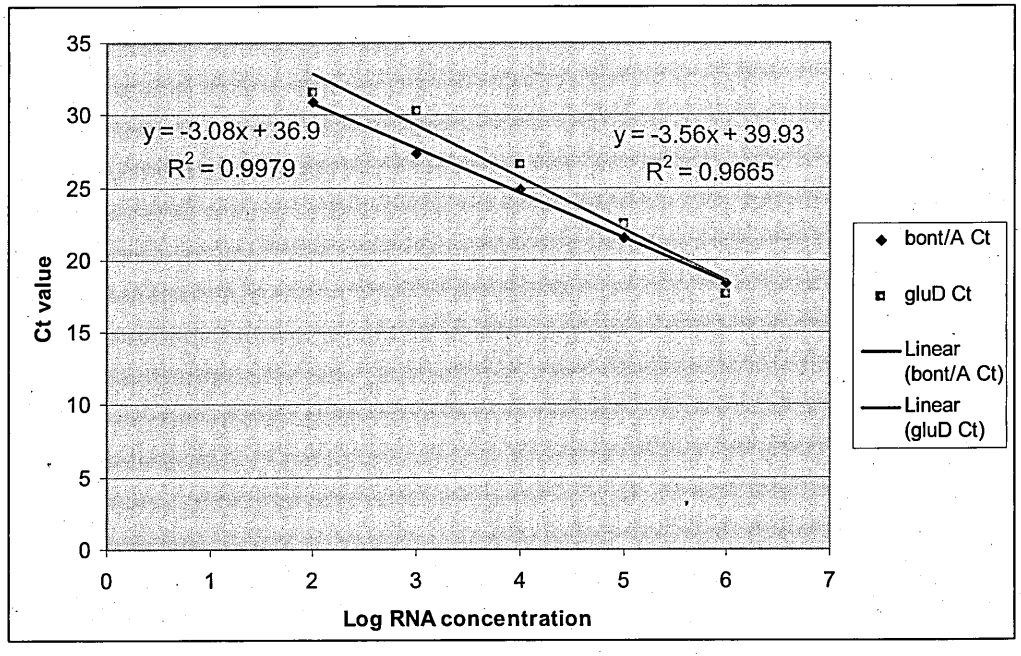


Figure 58: RT-qPCR of Serial dilution of total ATCC 19397 RNA. The slope of the line allows the calculation of efficiency.

Table 39: Calculation of relative gene expression from crossing threshold (Ct) values obtained from RT-qPCR experiments.

Replicate A

	Ct	Ct	Ct - Ct at 7h	Ct - Ct at 7h	Target efficiency ^ Ct - Ct at 7h		
Time	bont/A	gluD	bont/A	gluD	bont/A	gluD	RGE
7	21.5	17.7	0	0	1	1	1
8	18.8	16.7	2.7	0.9	7.3	1.8	3.9
9	17.8	16.6	3.6	1.1	15	2	7.6
10	18.4	16.9	3.1	0.7	9.8	1.6	6.1
11	18.5	16.4	2.9	1.3	8.8	2.3	3.8
12	19.3	16.2	2.2	1.5	5.2	2.6	2
13	19.7	15.9	1.8	1.8	3.8	3.2	1.2
24	28.6	21.8	-7.2	-4.1	0	0.1	0.1

Replicate B

	Ct	Ct	Ct - Ct at 7h	Ct - Ct at 7h	Target efficiency ^ Ct - Ct at 7h		
Time	bont/A	gluD	bont/A	gluD	bont/A	gluD	RGE
7	21.1	18	0	0	1	1	1
8	19.2	17.3	1.9	0.7	4.2	1.5	2.8
9	18.4	17.3	2.7	0.6	7.5	1.5	5
10	18.1	17.2	3	0.7	9.7	1.6	6
11	19.2	16.7	1.9	1.3	4.2	2.3	1.9
12	19.4	16.5	1.7	1.4	3.6	2.3	1.6
13	18.9	16.2	2.2	1.8	5.4	3.1	1.7
24	27.9	21.8	-6.7	-3.8	0	0.1	0.1

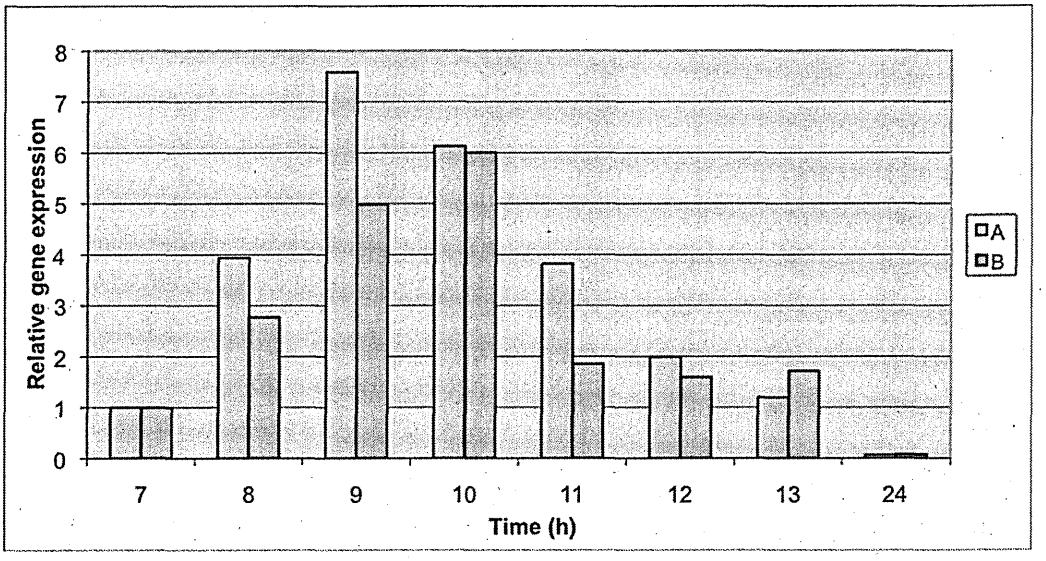


Figure 59: Relative gene expression of *bont/A* compared to *gluD* from 7 to 13 hours in two biological replicates A and B. Relative expression peaks at 9 h in sample A and 10 h in sample B before decreasing as the cells enter stationary phase. Based on this data three time points were selected for analysis from each replicate, from replicate A 7 h, 9 h and 13 h were selected and from replicate B 7 h, 10 h and 13 h were selected.

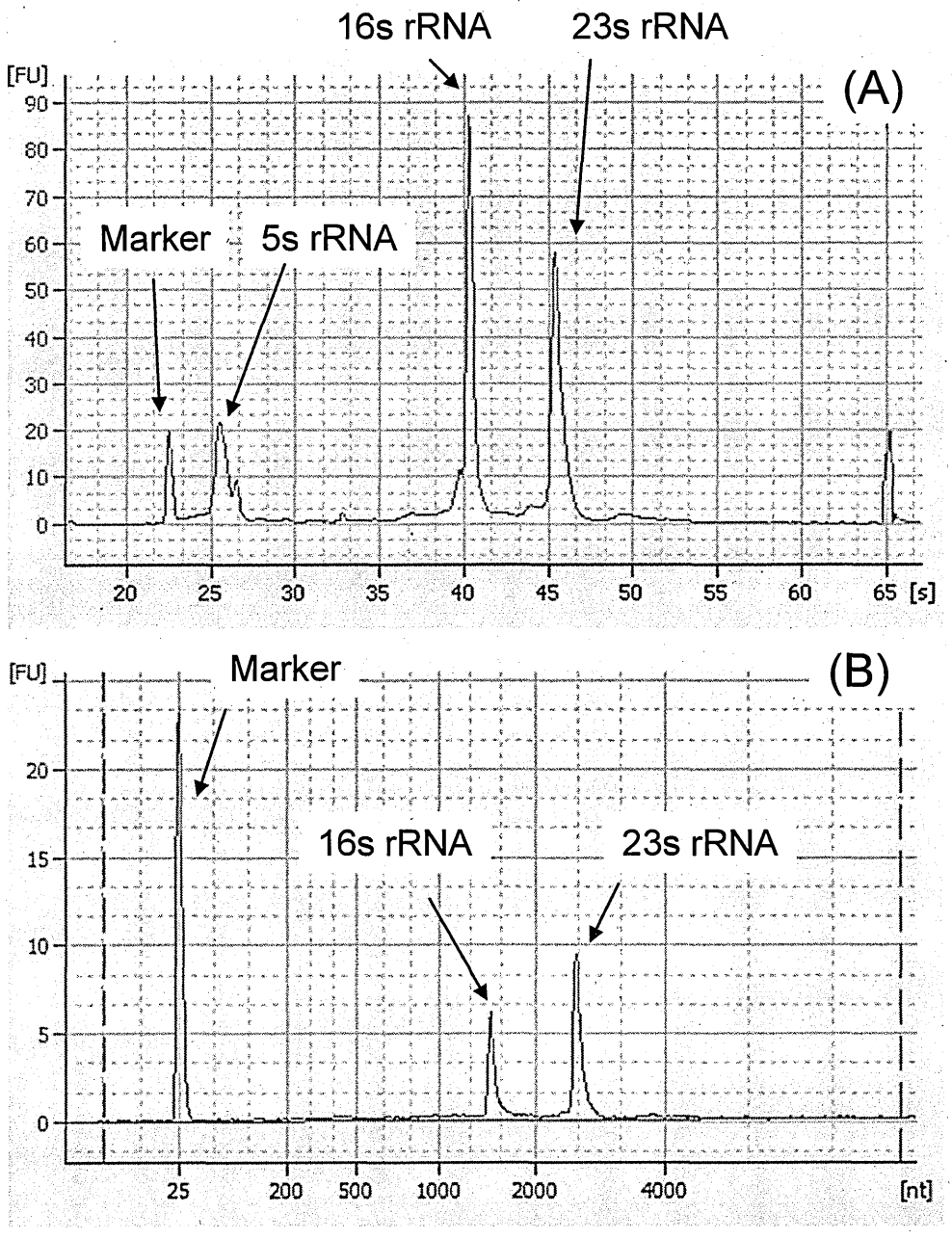


Figure 60: Presence of rRNA (5S, 16S and 23S) in extracted RNA before (A) and after (B) reduction of rRNA using Ribo-Zero rRNA reduction kit (Gram positive bacteria).

3.4.3. Analysis of gene expression data from the RNA-seq dataset

In total, 12 samples were analysed by RNA-seq. These samples were multiplexed on one flow cell of a SOLiD 4 machine. These 12 samples consisted of two technical replicates of two biological replicates of three time points.

The resulting sequencing reads were then mapped against the published genome of *C. botulinum* A1 19397. The number of RNA sequencing reads mapped to each of the 3700 coding sequences annotated in the *C. botulinum* A1 ATCC 19397 was determined. This measure was used to calculate the Reads Per Kilobase of gene per Million mapped reads (RPKM), i.e. the number of reads which map to a gene normalised for the total number of reads mapped to a sample and the length of the gene. This allows comparison of gene expression between samples with different sequencing coverages and between genes of different lengths, it can be thought of as the concentration of a transcript. The second approach used for 'read count data' is differential gene expression (DGE) analysis. Most methods for calling differentially expressed genes use untransformed data i.e. read counts rather than RPKM values. The DGE analysis package used in this work is edgeR, an R package (http://www.r-project.org/) that calls differentially expressed genes using a negative binomial model to account for biological variation (Oshlack et al., 2010; Robinson et al., 2010).

3.4.3.1. C. botulinum A1 ATCC 19397 gene expression

On average, more than 40 million reads were obtained for each of the 12 samples. When these reads were mapped to the reference genome an average of 4.5 million or 11.2% (standard deviation = 1.3 million or 3.3%) of the reads mapped

unambiguously per sample. A significant proportion of reads mapped to rRNA encoding regions, with an average 16.1% of mid-log, 20.5% of late log and 30.1% of early stationary reads mapping to 5S, 16S and 23S rRNA.

There were 3700 annotated coding sequences (CDS) in the reference genome. When technical and biological replicates for each time point were combined there were 2684 CDSs at mid-log, 2697 CDSs at late-log and 2326 CDSs at early stationary phase with RPKM greater than 10. This threshold was set heuristically and is not intended to convey biological meaning but rather to differentiate between signal and noise. The CDSs transcribed at each time-point were compared. There was a large core transcriptome of 2208 transcripts present at all three time-points. There were also 643 transcripts that were only present at one or two of the time-points (see Figure 61 for details).

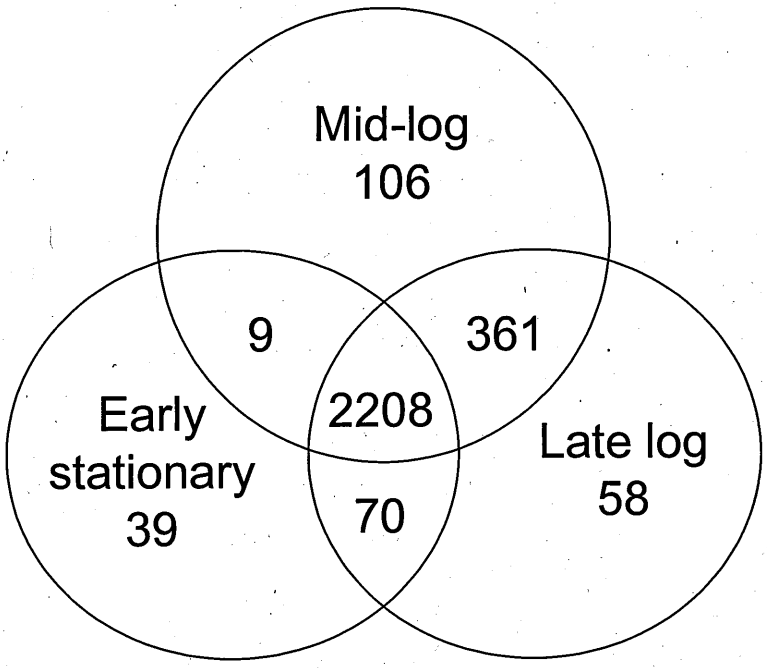


Figure 61: Venn diagram showing distribution of CDSs with RPKM greater than 10.

3.4.3.2. Comparison of RNA-seq gene expression data with microarray data

The RNA-seq results were validated against data from a microarray time-course analysis of proteolytic *C. botulinum* (Artin et al., 2010). The Artin study was performed using *C. botulinum* A1 ATCC 3502 which has nearly identical genome synteny with *C. botulinum* A1 ATCC 19397 (Hill et al., 2009). Orthologous loci were identified by an all vs all BLASTp approach and normalised microarray expression results compared with normalised RNA-seq results (RPKM) for each ortholog. The correlation co-efficent (r^2) between microarray and RNA-seq data was found to be 0.37 (Figure 62).

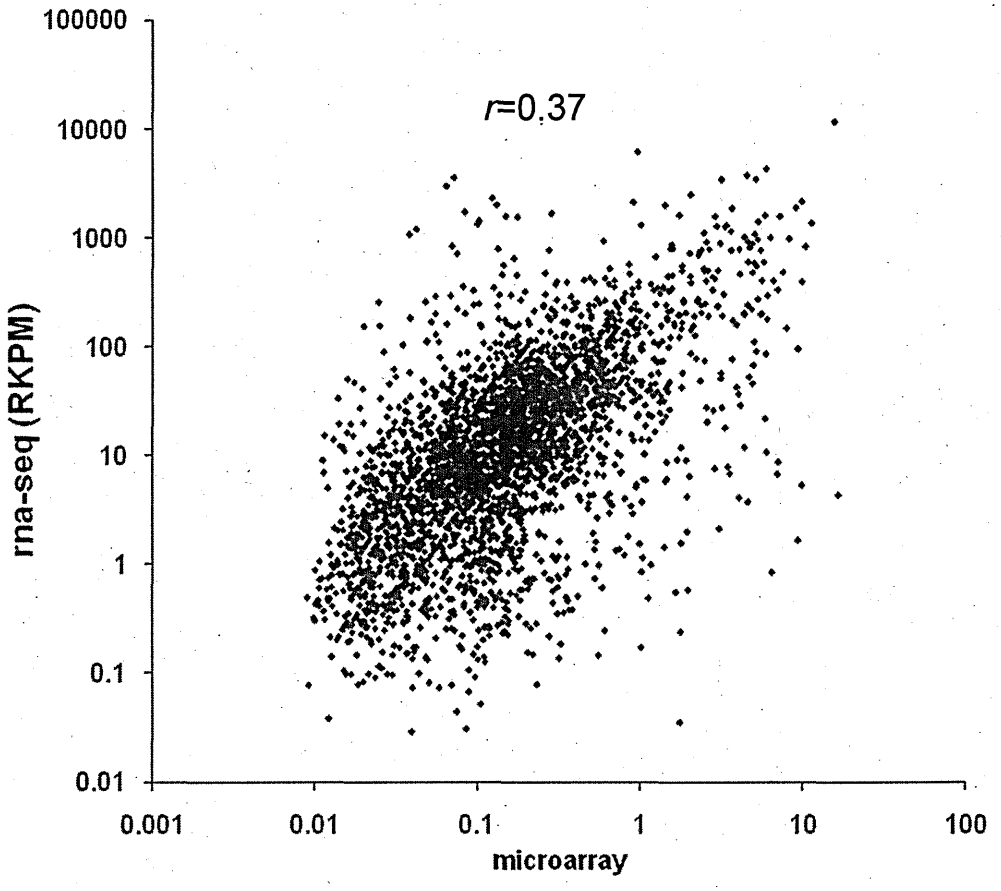


Figure 62: Comparison of gene expression results of microarray study (Artin et al., 2010) with RNA-seq data obtained in this study, samples taken at mid-log phase.

3.4.3.3. Genes which showed differential expression between timepoints

Differential gene expression analysis was performed on the RNA-seq data using the package edgeR implemented in the statistical programming language R. There were 252 CDSs (6.8% of total CDSs) that showed significant differential expression (FDR < 0.01) between mid-log and late log and 575 CDSs (19.7% of total) that showed significant differential expression between late-log and early stationary phases (Table 40). There were 93 CDSs that showed differential expression at both time points (Figure 63).

Differentially expressed genes were investigated for a potential role in pathogenicity in Results section 3.4.3.7.

Table 40: Number of genes showing significant (false discovery rate, FDR <0.01) positive or negative fold change between mid-log and late-log or between late log and early stationary phases.

	Mid Log vs Late Log	Late Log vs Early Stationary
Positive fold change	139	305
Negative fold change	110	270
Total	249	575

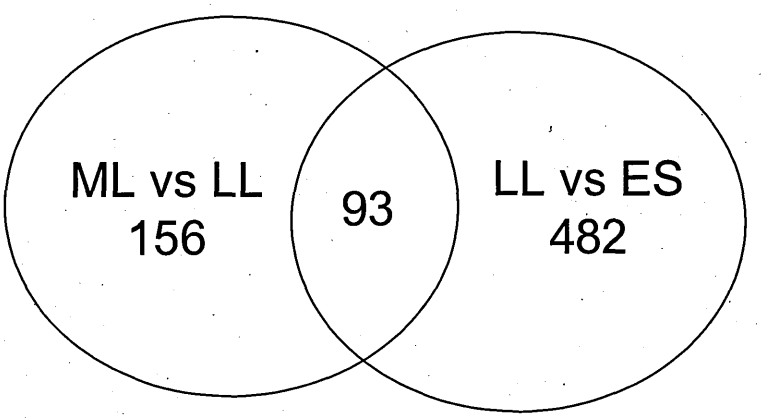


Figure 63: Venn diagram showing the number of genes which were differentially expressed (FDR < 0.01) between mid-log (ML) and late log (LL) and between late log and early stationary (ES) phase.

3.4.3.4. Expression of the botulinum toxin cluster genes

Expression from the genes comprising the *bont/A* cluster and *cloSI* (clostripain) was examined (Table 41,); *botA, ntnH, ha33, ha17* and *ha70* were actively expressed at mid-log stage (395th to 531st most highly expressed transcripts when CDSs ranked by RPKM). Visual inspection indicates transcription of the *ntnH-botA* operon and the *ha33-ha17-ha70* operon in contiguous, separate transcripts at all three time-points (Figure 65). *botR* expression was very low, only just above the noise threshold (RPKM of 10 or less was deemed noise) and among the genes with the lowest expression (2650th of 2684 CDSs with an RPKM of > 10). *botR* expression was 3.6% that of *botA*. At mid-log *cloSI* was the 290th most expressed gene, showing higher expression than any BoNT complex gene (Table 41 (A) and (B)).

Between mid-log and late log *botA* and *ntnH* expression increased 5.13 and 5.45 fold respectively while *ha33, ha17* and *ha70* increased by 4.22, 4.77 and 6.03 fold respectively (all false discovery rate (FDR) <0.01). At late log phase these genes were the 78th to 134th most highly expressed genes (Table 41). *botR* expression showed no significant change and remained amongst the least expressed genes, at 1.5% of *botA* expression. *cloSI* expression decreased 4.37 fold between mid-log and late log (FDR = 4.4E-03) becoming the 862nd most expressed gene at late log (Table 41 (A) and (B)).

There was a further increase in expression of *botA, ntnH, ha33, ha17* and *ha70* between late log and early stationary by 1.31, 1.27, 1.60, 1.66 and 1.64 fold respectively. However, these changes were not deemed to be statistically significant (FDR <0.01) by edgeR. These genes were the 40th to 94th most

expressed genes at early stationary phase. *botR* and *cloSI* also showed no significant change between log and early stationary phases (Table 41 (A) and (B)). *botR* was expressed at 0.6% the level of *botA*.

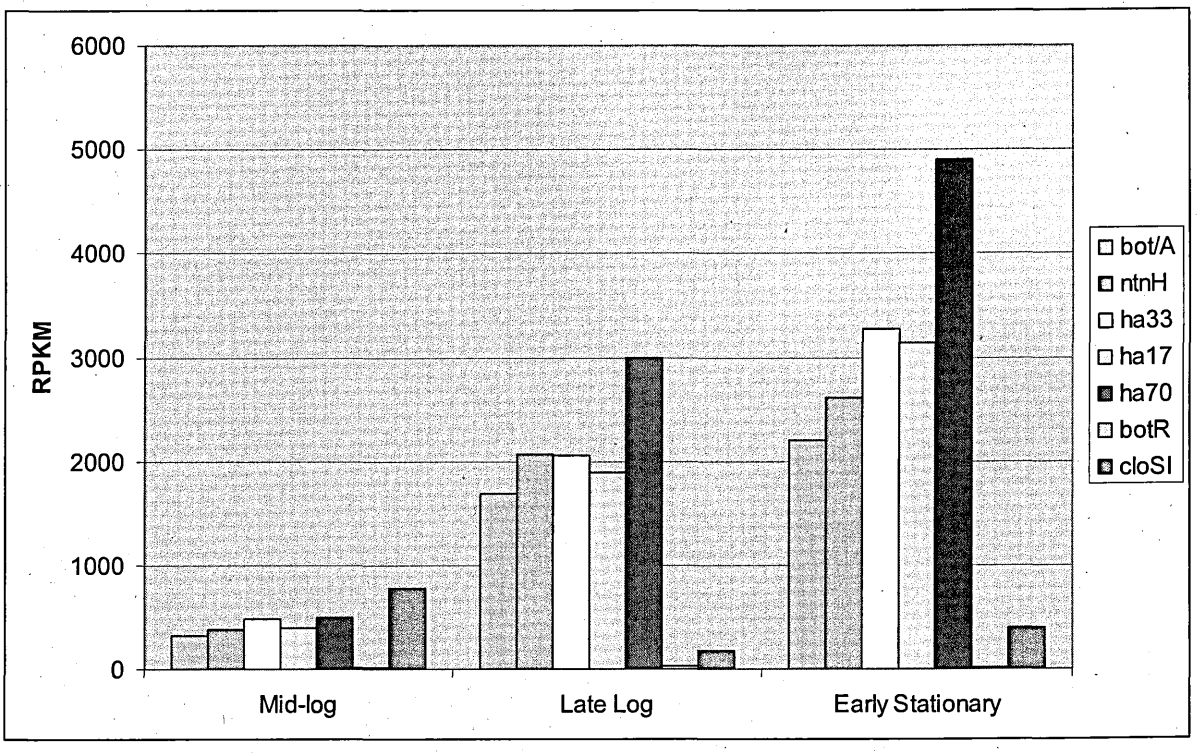


Figure 64: Expression of *bont/A* cluster genes at mid-log (ML), late log (LL) and early stationary (ES) phases.

Table 41: Expression of *bont/A* cluster genes at mid-log (ML), late log (LL) and early stationary (ES) phases. (A) Shows normalised transcript level, RPKM, of the toxin gene cluster and clostripain and the rank of each gene compared to the expression levels for all the genes at that time point (B) shows log fold change and associated confidence scores as calculated by edgeR.

(A)

Gene	RPKM ML	ML rank	RPKM LL	LL rank	RPKM ES	ES rank
botA	328.01	531	1681.75	134	2199.01	94
ntnH	378.95	482	2066.12	116	2617.17	79
ha33	486.04	402	2053.29	117	3277.06	65
ha17	397.25	464	1893.12	128	3137.17	69
ha70	497.33	395	2996.57	78	4904.85	40
botR	11.76	2650	25.30	2202	15.26	2,070
cloSI (clostripain)	774.04	290	177.21	862	396.41	327

(B)

Gene	Fold change ML to LL	FDR	Fold change LL to ES	FDR
botA	5.13	4.40E-04	1.31	7.14E-01
ntnH	5.45	3.40E-04	1.27	7.12E-01
ha33	4.22	3.49E-03	1.6	4.24E-01
ha17	4.77	8.86E-04	1.66	4.37E-01
ha70	6.03	1.13E-04	1.64	4.31E-01
botR	2.15	1.54E-01	-1.66	5.37E-01
***cloSI* (clostripain)**	-4.37	4.41E-03	2.24	2.29E-01

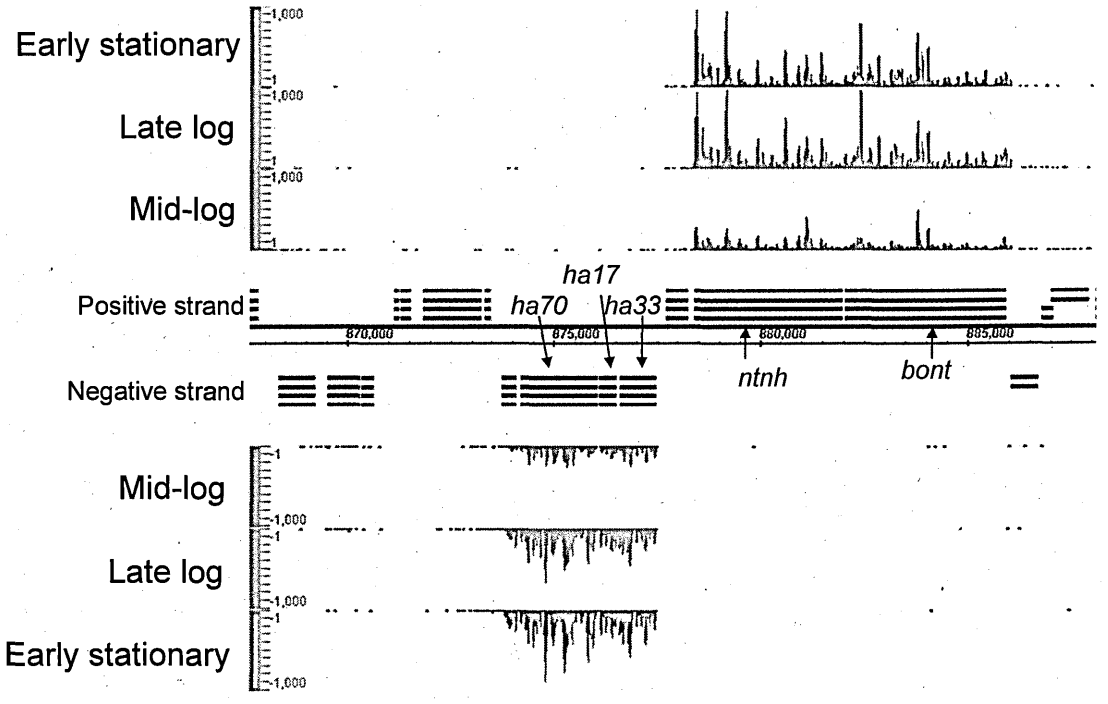


Figure 65: Expression from *bont* gene cluster at mid-log, late log and early stationary phases. Y axis is read depth.

3.4.3.5. Comparison of *bont/A* expression as measured by RT-qPCR relative to *gluD* and as measured by RNA-seq

The *botA* expression values obtained by RNA-seq were compared to those obtained from the RT-qPCR results (Figure 66). The expression of *botA*, measured by RT-qPCR relative to *gluD*, peaked at late log phase before decreasing in early stationary phase. This conclusion was not supported by the measured expression of the same gene using RNA-seq. The results from the latter showed a similar increase between mid-log and late log but then showed a small, insignificant increase between late log and early stationary phase. In order to resolve this contradiction, expression levels of *gluD* were examined in the RNA-seq data. There was an insignificant (FDR >0.01) increase in expression of 1.3 fold between mid-log and then a significant increase of 5.8 fold (FDR = 2.8×10^{-12}) between late log and early stationary. This would result in the relative expression of *bont/A* decreasing at early stationary phase.

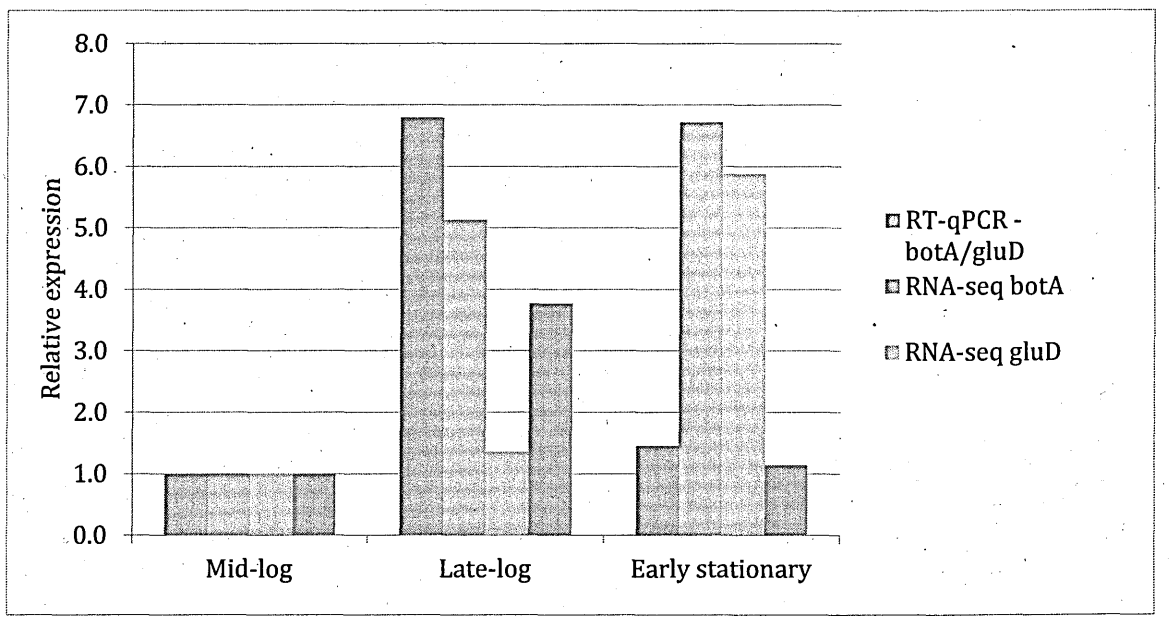


Figure 66: Comparison of RT-qPCR relative gene expression and RNA-seq exression results. Expression of *botA* measured by RT-qPCR relative to *gluD*, *botA* expression measured by RNA-seq and *gluD* expression measured by qPCR.

3.4.3.6. Identifying genes which are suitable reference genes for relative gene expression experiments

The RNA-seq dataset was examined to determine genes that showed stable expression across biological replicates and time points. To this end the coefficient of variance was calculated (standard deviation divided by the mean) from the six RPKM values for each gene (three time points, two biological replicates). The expression details of the 10 genes with the lowest co-efficient of variation are given in Table 42 with *gluD*, used as a reference gene in section 3.4.2, included for comparison (only genes with an average RPKM > 100 were included). The gene with the lowest coefficient of variation was *clb_0260*, which encodes a lipoprotein. Other genes with stable expression across time points and biological replicates were *spoVG*, which encodes a sporulation protein; *rpoD* that encodes an RNA polymerase sigma factor and *buk_2* which encodes a butyrate kinase. The genes included in the MLST scheme for an organism are often considered as candidates for reference genes in relative gene expression results. For *C. botulinum* the 7 MLST genes are *mdh, aceK, rpoB, aroE, hsp60, oppB* and *recA* (Jacobsen et al., 2008). The expression of the *C. botulinum* MLST scheme genes across the three time points and two biological replicates was also examined (Table 44) – *recA* was the MLST gene with the lowest coefficient of variation of 0.22, making it the 28th most stable gene expressed by ATCC 19397. The function of the 10 most stably expressed genes is described in Table 43.

Table 42: Expression levels (RPKM) of the 10 genes with the lowest co-efficient of variance across both biological replicates and all timepoints. *gluD*, used as the reference gene in section 3.4.2 is included. T1B1 to T3B2 are RPKM values at time point 1, 2 and 3 for biological replicates 1 and 2.

Feature ID	T1B1	T1B2	T2B1	T2B2	T3B1	T3B2	Average	Standard deviation	Coefficient of variation
CLB_0260	1,334	995	1,049	1,193	1,125	1,198	1148.9	120.5	0.1
spoVG	4,727	3,916	4,608	5,321	4,499	5,293	4727.3	528.8	0.11
CLB_2031	160	152	160	189	123	176	159.9	22.2	0.14
rpoD	243	213	239	257	169	215	222.7	31.1	0.14
buk_2	200	173	206	221	155	230	197.5	28.8	0.15
CLB_1337	430	292	432	437	355	446	398.7	62.1	0.16
CLB_2907	205	180	260	282	219	254	233.5	38.5	0.16
gcvPA	170	212	136	171	156	216	176.8	31.6	0.18
CLB_3662	179	251	213	242	152	187	204.1	38.3	0.19
CLB_3433	266	227	243	298	161	234	238.2	45.6	0.19
gluD	1167.1	1671.1	1546.1	2313.6	8689.8	13994	4896.9	5279.4	1.08

Table 43: Functional role of the 10 genes with the lowest coefficients of variance.

Gene	Main role	Sub role	Common name
CLB_0260	Cell envelope	Other	lipoprotein, bmp family
spoVG	Cellular processes	Sporulation and germination	stage V sporulation protein G
CLB_2031	Hypothetical protein	Hypothetical protein	hypothetical protein
rpoD	Transcription	Transcription factors	RNA polymerase sigma factor RpoD
buk-2	Energy metabolism	Fermentation	butyrate kinase
CLB_1337	Unknown function	Enzymes of unknown specificity	dinitrogenase iron-molybdenum cofactor family protein
CLB_2907	Unknown function	General	UBA/TS-N domain protein
gcvPA	Energy metabolism	Amino acids and amines	glycine cleavage system P protein, subunit 1
CLB_3662	Unknown function	Enzymes of unknown specificity	radical SAM domain protein
CLB_3433	Hypothetical proteins	Conserved	conserved hypothetical protein
gluD	Energy metabolism	Amino acids and amines	glutamate dehydrogenase, NAD-specific

Table 44: Expression levels (RPKM) of the 6 MLST scheme genes with an RPKM greater than 10 for at least one time-point were examined. T1B1 to T3B2 are RPKM values at time point 1, 2 and 3 for biological replicates 1 and 2.

Feature ID	T1B1	T1B2	T2B1	T2B2	T3B1	T3B2	Average	Standard deviation	Coefficien of variation
recA	510.0	306.2	366.9	417.3	272.5	383.5	376.1	84.1	0.22
CLB_1385 (oppB)	11.9	16.7	5.6	13.0	8.8	8.8	10.8	3.9	0.36
aroE	55.4	46.8	27.8	36.7	18.9	24.6	35.0	14.0	0.40
rpoB	955.8	1139.4	533.3	729.7	129.1	183.7	611.8	408.2	0.67
hsp60	266.1	117.1	104.3	83.7	31.0	26.1	104.7	87.5	0.83
CLB_3479 (aceK)	757.1	1039.8	1625.6	2427.6	7514.9	14521.5	4647.8	5439.3	1.17

3.4.3.7. Putative pathogenicity associated genes with increased expression in early stationary phase

Differential gene expression data was analysed to identify putative pathogenicity associated genes with significant fold changes. Fold changes with a False Discovery Rate (FDR) of less than 0.01 were deemed significant, in total 731 genes had fold changes supported by FDRs < 0.01.

Six thermolysin metallopeptidase genes (*npr-1* to *6*) showed significant up-regulation between late log and early stationary phase. These six genes are encoded in the same operon in the genome of ATCC 19397 (Figure 67) and show an average 79% +/- 6% similarity to each other on the nucleotide level. *npr-6* and *npr-5* show the highest level of expression with RPKMs of 349.9 and 238.2, increasing 4.4 fold and 4.6 fold between late log and early stationary phase respectively. *npr-4, npr-3* and *npr-2* show the second highest level of expression with RPKMs of 85.7, 89.0 and 83.1, increasing 3.4 fold, 3.5 fold and 2.8 fold respectively between late log and early stationary phase. *npr-1* has the lowest level of expression with an RPKM of 8.16, increasing 1.4 fold from late log to early stationary phase (Table 45).

C. botulinum encodes a gene, *clb_1609*, for a hemolysin-III protein which shows significant similarity (BLAST E-value = 1.6×10^{-57}) to a *Bacillus cereus* pore forming cytotoxin which disrupts hemocytes (Table 46). This protein shows strong up-regulation between late log and early stationary phase. The expression level of *clb_1609* increased 4.2 fold to an RPKM of 10695.2.

There was significant up-regulation of a bacteriocin/streptolysin-encoding gene, *clb_0527*, between late log and early stationary phase. There was a 1.6 fold increase to an RPKM of 3829. Bacteriocins are toxins produced by bacteria to inhibit the growth of similar or closely related bacterial strains (Table 46).

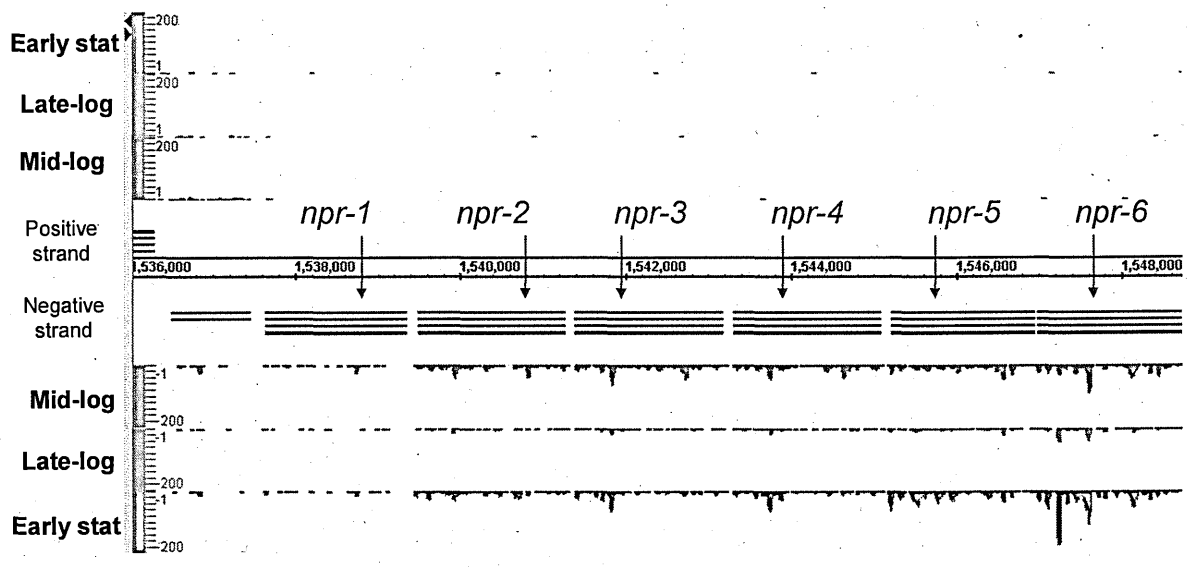


Figure 67: Expression of *npr1-6* at mid-log, late log and early stationary phases. The Y-axis scale is 200 reads. The average length of the *npr* genes is 1762 bp.

Table 45: Thermolysin metallopeptidase genes which show increased expression between late log and early stationary phase - fold change, False Discovery Rate (FDR) and expression level (RPKM)

Gene	Fold change - late log to early stat	FDR	RPKM
npr-1	1.42	9.10×10^{-3}	8.16
npr-2	3.51	4.60×10^{-10}	83.05
npr-3	2.75	3.61×10^{-8}	89.01
npr-4	3.38	6.82×10^{-10}	85.7
npr-5	4.58	2.76×10^{-10}	238.24
npr-6	4.4	1.21×10^{-9}	349.91

Table 46: Putative pathogenesis associated genes which show increased expression between late log and early stationary phases - fold change, False Discovery Rate (FDR) and expression level (RPKM)

Gene	Fold change - late log to early stat	FDR	RPKM
CLB_1609 (hemolysin)	4.18	9.00×10^{-8}	10695
CLB_0527 (bacteriocin)	1.67	1.50×10^{-3}	3829.1

3.4.4. Linking the transcriptome and proteome data

Transcription and translation do not always correlate due to the diverse array of post-transcriptional phenomena influencing the rate at which mRNA is translated into protein (Maier et al., 2009). The expression of genes encoding proteins that were identified in the culture supernatant at 24 h and 96 h was examined. However, the complexity of the relationship between mRNA concentration and protein concentration necessitates caution when analysing this data.

The protein with the lowest RPKM that was identified in the culture supernatant was a putative NADPH-dependent FMN reductase (A7FSA3) involved in electron transport. The gene encoding this protein had a maximum RPKM at the timepoints measured of just 7. Other proteins with low RPKMs were involved in a phosphotransferase system (A7FVJ3, max RPKM = 13.5), a glyoxalase family protein (A7FU43, max RPKM = 37) and a MarR family transcriptional regulator (A7FSQ9, max RPKM = 75). There were six proteins detected by LC-MS/MS for which there was no RNA expression at any time point. These proteins were all involved in energy metabolism, with five of the six involved in the utilisation of amino acids for energy (A7FPY4, a 2-hydroxyglutaryl-CoA dehydratase; A7FW29 and A7FW44, D-proline reductases; A7FYN7 and A7FZI3, CoA transferases) and one involved in fermentation (A7FZ18, butyrate kinase).

The average expression of genes encoding proteins present at 24 h, 96 h and both time-points was calculated for comparison. At mid-log, late log and early stationary phases the median gene expression of supernatant proteins present at 24 h and 96 h was higher than the median gene expression of proteins present at only 24 h or 96 h. At early stationary phase, the expression of genes encoding

proteins present at only 96 h is lower than the expression of genes encoding proteins present at only 24 h (Table 47).

The results of the BoNT/A endopeptidase assay (section 3.2.3) were compared with the *bont/A* expression results from the RNA-seq experiment. Between mid-log and early stationary phase, the *bont/A* RPKM increased 6.7 fold. Over the same time period BoNT/A, as measured by endopeptidase assay increased only 2.1 fold. BoNT/A concentration peaked at 24 h, increasing 4.8 fold between 7 h and 24 h.

Table 47: Median expression of genes (RPKM) encoding proteins detected at only 24 h, only 96 h or both at mid-log, late-log and early stationary phase.

	Mid-log	Late log	Early stationary
24 h	663.0	969.3	517.5
96 h	386.0	640.0	96.5
24 h and 96 h	1787.8	1617.3	751.5

3.4.5. Small RNAs in global gene expression

The different expression patterns of the *bont/A* transcript and the concentration of the BoNT/A protein raises the possibility of the post-transcriptional regulation of *bont/A*. One major mode of post-transcriptional regulation is by the interaction of small RNAs (sRNA) with mRNA. Here the sRNA complement of *C. botulinum* will be examined.

Transcription from the non-coding regions of the *bont/A* gene cluster was examined for the presence of possible sRNAs by visual inspection. The sequences of any regions of transcription were then compared with the Rfam functional RNA database (http://rfam.sanger.ac.uk/). There was no significant non-mRNA transcription in the region of the *bont/A* gene cluster.

There could be an sRNA encoded elsewhere on the genome which interacts with the *bont/A* mRNA. A thorough characterisation of sRNAs from across the *C. botulinum* ATCC 19397 genome was carried out. All sRNAs identified were examined for homology with the *bont/A* mRNA which could indicate that the sRNA played a role in the post-transcriptional regulation of *bont/A*. No candidates for interaction with the toxin gene were identified but an active transcriptional landscape of non-coding sRNAs was uncovered.

There were 17 transcriptionally active T-boxes identified (Table 48). T-boxes are sRNAs that are located in the 5' untranslated region (5'UTR), upstream of mRNA encoding amino acid-tRNA ligases. One particularly interesting example is that of *valS* (a valine-tRNA ligase) and CLB_3200 (a GNAT family acetyltransferase) that appear to be co-transcribed and jointly regulated by a single T-box (Figure 68).

There were also 19 transcriptionally active riboswtiches identified in the *C. botulinum* genome (Table 49). Examples of riboswitches identified in *C. botulinum* include magnesium riboswitches, thioamine pyrophosphate riboswitches and purines/pyrimindes riboswitches among others.

Table 48: Genes which had transcriptionally active T-boxes or S-boxes identified in their upstream region.

5' UTR mechanism	Gene symbol/locus	Gene function
T-box	serS-1	Serine-tRNA ligase
T-box	CLB_1163	Class V aminotransferase
T-box	aspC	Aspartate transaminase (changes aspartate to glutamate)
T-box	metG	Methionine-tRNA ligase
T-box	leuS	Leucine-tRNA ligase
T-box	thrS	Threonine--tRNA ligase
T-box	valS *	Valine-tRNA ligase
T-box	CLB_3200 *	Acetyltransferase, GNAT family
T-box	ileS	Isoleucine-tRNA ligase
T-box	CLB_0276	Sodium:dicarboxylate symporter family protein
T-box	serS-2	Serine-tRNA ligase
T-box	argS	Arginine-tRNA ligase
T-box	alaS	Alanine-tRNA ligase
T-box	tyrS	Tyrosine-tRNA ligase
T-box	CLB_2881	Putative membrane protein
T-box	pheS	Phenylalanine-tRNA ligase
T-box	CLB_0768	Conserved hypothetical protein, potential methyltransferase activity
S-box	metK	S-adenosylmethionine synthetase
S-box	metF	Methylenetetrahydrofolate reductase
S-box	CLB_3459	Na+/H+ antiporter family protein

Table 49: Cis-encoded Regualtory RNA families identified by manual analysis of *C. botulinum* transcriptome

RNA family	Summary of role	Number identified	Examples of target genes
T-box	Up-regulates aminoacyl-tRNA synthetase genes in the presence of uncharged tRNA	17	*leuS*, *thrS*, CLB_3200-*valS* operon
S-box (aka SAM-I riboswitch)	When SAM-1 riboswitch is bound by S-adenosylmethioine the downstream genes which are typically involved in methioine synthesis are down regulated.	3	*MetK*, *MetF*
M-box (*ykoK* leader)	Transcription is down regulated in by the presence of magnesium of other divalent ions, target genes typically involved in magnesium homeostasis.	2	*mgtA*,
Ribosomal protein leader	When cellular concentrations of a ribosomal protein are too high, the protein will bind to this region and down regulate expression of the ribosomal protein gene.	2	IF-3 operon, rplJ
PyrR binding site	Regulates a variety of genes involved in pyrimidine biosynthesis and transport	1	codB
GlmS ribozyme	A dual ribozyme-riboswitch, when the riboswitch moiety is bound by glucosamine-6-phosphate the ribozyme moiety is activated and self cleavage occurs leaving a 5' hydroxyl group which targets the transcript for degredation by RNase J1.	1	*glmS*
Lysine riboswitch	Regulates a variety of genes in a lysine dependeny manner, typically those involving lysine metabolism.	1	CLB_0454, an Na+/H+ antiporter family protein
Thioamine pyrophosphate riboswitch	Binds to thiamine pyrophosphate and regulates a large number of genes	3	Thiamine transporter, purF-2
Glycine riboswitch	Consists of two metabolite binding domains which act in tandem to inititate expression of genes involved in glycine degredation	2	Glycine degredation operon, *grdX*, *B*, *D*, *A* and *C*
Purine riboswitch	Selectively recognises purines, controls expression of genes involved in purine synthesis and transport	3	CLB_2127, purine transporter, purine synthesis genes
Flavin mononucleotide riboswtich	Recognises flavin mononucleotides, controls expression of genes involved in FMN synthesis and transport	1	*rib* operon, riboflavin biosynthesis protein

There were also some noteworthy novel features of unknown function. There were multiple examples of short areas of sense-antisense transcription (Figure 69). Possible toxin-antitoxin systems account for a fraction of the sense-antisense transcriptional activity – any function the other sites may have remains obscure. An sRNA which showed strong growth phase regulation was identified, the RPKM at mid-log was 551, at late log it was 851 and at late stationary it was 36397. This dramatic increase in expression was not mirrored in either of the flanking genes, which are both on the positive strand while the sRNA is on the negative strand.

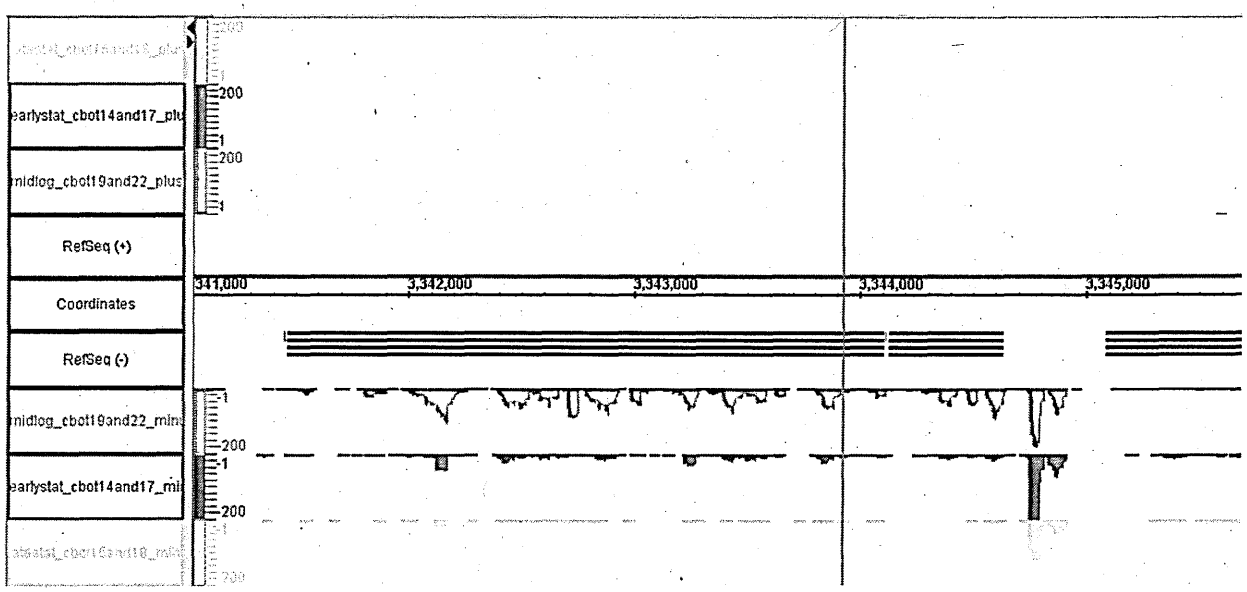


Figure 68: Transcriptional activity of a T-box region. The *valS*-CLB_3200 operon can be seen downstream of the T-box. CLB_3200 is the shorter gene which immediately follows the T-box and *valS*, a valine synthesis gene which would be expected gene to be associated with the T-box is the larger gene and second in the operon.

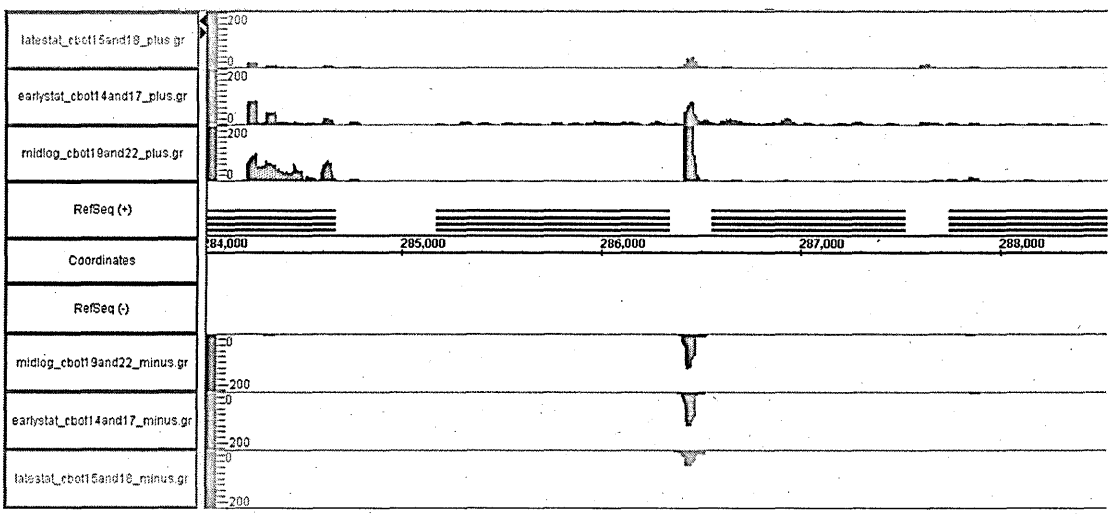


Figure 69: An example of sense-antisense transcription from intergenic region.

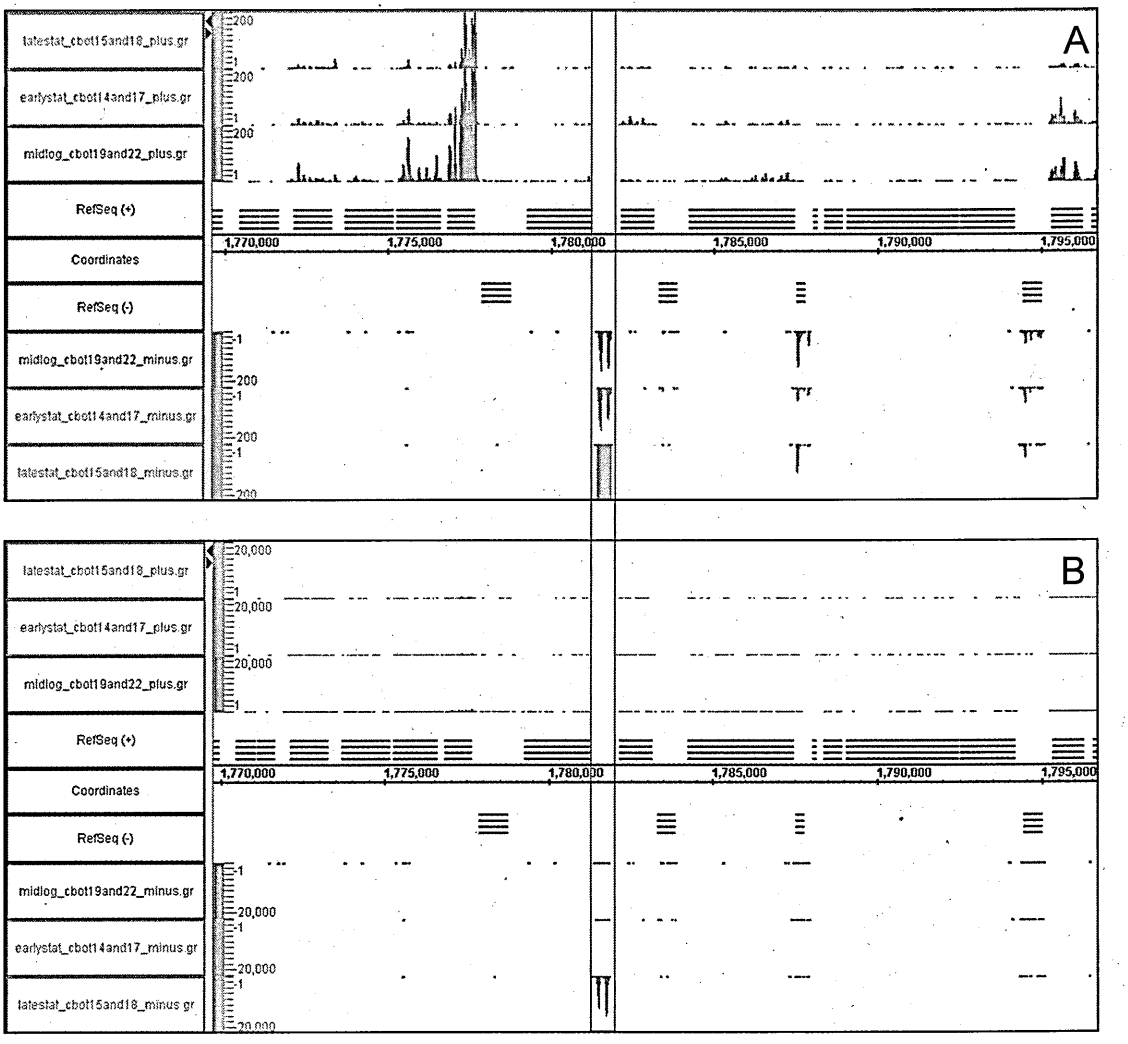


Figure 70: Level of expression in the conserved region to which the sRNA (highlighted in purple box) belongs (A) shows expression of the growth phase regulated sRNA at a read depth of 200 allowing the visualisation of transcription from other areas within the conserved region (B) visualisation of the same conserved region as (A) but the scale is set at 20000 read depth showing the extremely high level of transcription of the sRNA at early stationary phase compared to early growth phases and to the rest of the conserved region.

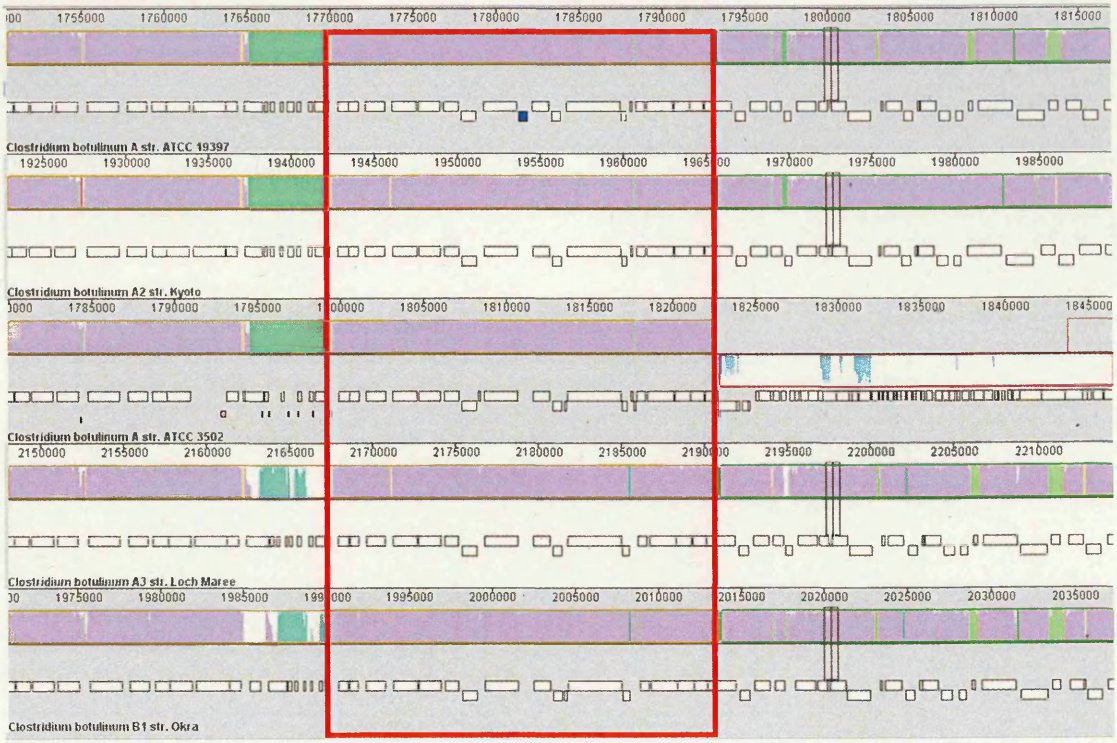


Figure 71: The presence of a growth phase regulated sRNA (dark blue, top centre) in a large region (highlighted in red box, 23 kbp) of conservation among proteolytic *C. botulinum*. Conserved genes upstream of the sRNA include an alanine racemase involved in cell wall synthesis, 2 methyl accepting chemotaxis proteins, a xanthine/uracil permease family protein and a short chain oxidoreductase. Downstream there is a Snf2/Rad54 family helicase, formylmethanofuran dehydrogenase E related protein, 3 peptide-opine-nichel uptake ABC transporter protein. Only proteins within this region are on the same strand as the sRNA (the oxidoreductase and 2 hypothetical proteins).

Table 50: Similarity of nucleotide sequence of growth phase regulated sRNA among proteolytic *C. botulinum* calculated using BLAST

Genome	Query coverage	Identity
Clostridium botulinum A2 str. Kyoto	100%	100%
Clostridium botulinum A str. Hall	100%	100%
Clostridium botulinum A str. ATCC 19397	100%	100%
Clostridium botulinum A str. ATCC 3502	100%	100%
Clostridium botulinum H04402065	100%	99%
Clostridium botulinum A3 str. Loch Maree	100%	98%
Clostridium botulinum B1 str. Okra	100%	98%
Clostridium botulinum F str. 230613	100%	97%
Clostridium botulinum F str. Langeland	100%	97%
Clostridium botulinum Ba4 str. 657	99%	97%
Clostridium botulinum A1(B) 2916	N/A	N/A
Clostridium botulinum Bf	N/A	N/A

3.4.6. Summary of findings of transcriptome investigation of *C. botulinum*

- *gluD* is not a suitable reference gene for relative gene expression experiments. Other, more suitable candidates include *rpoD* that encodes a transcription factor.
- The RNA-seq data shows that while expression of the toxin complex genes increases, expression of *botR* shows no significant change.
- There are genes that show up-regulation as the culture enters early stationary phase that encode proteins with homology to known virulence factors.
- There was an active transcriptional landscape of small RNAs. Some of these sRNAs had known functions in regulation cell metabolism or other cellular functions. There were also other, uncharacterised sRNAs including one that showed massive up-regulation as the culture entered stationary phase.

4. Discussion

4.1. In silico investigation of the botulinum toxin complex and C. botulinum proteome

The sequences of botulinum toxin and the associated non-toxin proteins were analysed using BLAST to identify similarity to other proteins and patterns of similarity between different types and subtypes of toxin cluster. It was discovered that there were multiple heterologous homologues of the OrfX and P-47 proteins in numerous species. These homologues were genomically co-localised and in 5 cases clustered with a putative toxin encoding gene.

4.1.1. In silico investigation of botulinum neurotoxin and non-toxic non-haemaglutinin

The amino acid sequences of BoNT and NTNH are much more similar than would be expected by chance, although still with 55% identity over 32% of the sequence. Interestingly, while the sequences are divergent the crystal structures of the two proteins are very similar, with a root mean square deviation of 2.3 Angstroms (Gu et al., 2012) and they are thought to have evolved from a common ancestral zinc binding protein via gene duplication (Inui et al., 2012). This shows that there is a long evolutionary history of functionally significant rearrangements in the BoNT encoding gene cluster.

Comparing the protein sequence dendrograms for BoNT and NTNH suggests that homologous recombination has resulted in novel arrangements within the BoNT encoding gene cluster, as previously suggested (Hill et al., 2009). All the BoNT/A

subtypes cluster together, with BoNT/B clustering separately. However, their corresponding NTNH sequences show a different pattern, with NTNH/A1 and NTNH/A5 clustering between between NTNH/A2/A3/A4/F and NTNH/B. This is thought to be due to a recombination event occurring approximately 1965 bp into the 3594 bp *ntnH* (Hill et al., 2009). It has been hypothesised that this recombination led to the combination of BoNT/A and the HA toxin complex, which was previously associated only with BoNT/B (Hill et al., 2009). This combination is most frequently associated with human disease. The *botR* sequences follow the same pattern as the NTNH sequences, indicating that BotR and NTNH share an evolutionary history, as expected from their genomic co-localisation. The fact that Lee et al. (2013) found that NTNH is the major interface between the BoNT-NTNH and HA sub-complexes may explain why NTNH is more conserved between different types of BoNT than BoNT itself. While the BoNT has undergone changes that have altered that neuronal cell binding and cleavage targets, NTNH has to conserve binding to the HA sub-complex or else compromise translocation across the intestinal epithelium (Lee et al., 2013). It has also been found that the NTNH encoded alongside BoNT/A1, B, C, D and G has a conserved motif that is responsible for the main interaction between the BoNT-NTNH and the HA sub-complex. This motif is missing from the NTNHs associated with BoNT/A2, E and F, which do not interact with HA (Lee et al., 2013; Lin et al., 2010).

Evidence of further recombination within the BoNT gene cluster is that while BoNT/A and BoNT/F are clearly divergent, with BoNT/F showing closer similarity to BoNT/E, NTNH/F clusters with the NTNH/A sequences from the OrfX encoding strains (e.g. NTNH/A3). This discrepancy indicates that there has been a horizontal gene transfer event between the BoNT/F gene cluster and an OrfX encoding BoNT cluster, resulting in the transfer of NTNH, as has been previously

noted (Raphael et al., 2010). This analysis extends previous work by showing that the OrfX proteins in the BoNT/F toxin gene cluster show greater similarity to the OrfX proteins of BoNT/A strains rather than BoNT/E strains.

Recent work has shown that a strain encoding BoNT/F6 evolved by successive disruption of two different ancestral precursors, indicating the role of recombination hotspots in the generation of diversity within the toxin gene cluster of type F strains (Carter et al., 2013). This extensive recombination may be one of the drivers of the considerable diversity within the type F gene cluster (Raphael et al., 2010).

The BoNT and TeNT proteins show similarity in the light chain, which is expected, as both are responsible for the enzymatic cleavage of the neuronal cell neuroexocytosis apparatus. There is also sequence and structural similarity between BoNT and TeNT in the amino terminus of the heavy chain that is responsible for the pH dependent translocation out of the endosome (Simpson, 2004). However, toward the carboxy terminus of the light chain the two proteins are highly divergent. It is this region that is responsible for binding to neuronal cells, and the divergence in this sequence contributes to their different *in vivo* targets and symptomatic effects. This analysis agrees with previous workers findings (Pellizzari et al., 1999).

4.1.2. *In silico* investigation of the haemaglutinin toxin complex

The amino acid sequence dendrograms for HA70, HA33 and HA17 (Figure 18, Figure 20, Figure 22) show the close relationship between the HA proteins found in type A and type B strains. This similarity is striking when compared with the low similarity between the BoNT/A and BoNT/B proteins. This indicates that a

horizontal gene transfer event has taken place resulting in the HA proteins being transferred between the two toxin clusters. It is likely that this transfer was part of the same event in which *ntnH* was transferred (Hill et al., 2009). Due to the fact that OrfX proteins have only been found associated with BoNT/A, while BoNT/A is found with both OrfX and HA proteins it is hypothesised that prior to this recombination, that BoNT/A was associated with OrfX (Hill et al., 2009). The identification here of a Ricin B lectin domain in HA33, which is also found in many other bacterial and plant toxins, has been previously reported (Arndt et al., 2005) thereby confirming the findings presented here.

Interproscan analysis of HA70 explained the significant match between HA70 and the *C. perfringens* enterotoxin; there was a *C. perfringens* enterotoxin domain in HA70. In *C. perfringens* enterotoxin, this domain is responsible for interaction with tight junctions in gut epithelium (Katahira et al., 1997). This is another piece of evidence that the HA sub-complex is facilitates interaction with the gut membrane. Interproscan analysis of HA17 also showed that it contains the common Ricin B lectin domain that is implicated in binding to sugars. This domain is frequently found in bacterial AB-toxins, among other proteins.

4.1.3. *In silico* investigation of the P-47 family toxin complex

There was a conserved domain in OrfX2, OrfX3 and P-47, known as the Clostridium P-47 protein domain (Pfam id IPR010567). The function of this domain is unknown. Due to the presence of this common domain in OrfX2, OrfX3 and P-47, these proteins will be referred to as P-47 family proteins. Whilst these results confirm and extend previous findings, there is also a novel insight into the P-47 family proteins. OrfX1 had no P-47 family domain, but did have one significant

BLAST match to a non-BoNT encoding species; a protein annotated as OrfX1 in *Paenibacillus dendritiformis*. The significance of this result is increased by the presence of homologues (>30% amino acid identity) of OrfX2, OrfX3 and P-47 in *P. dendritiformis*.

There were multiple proteins with significant BLAST matches to OrfX2, OrfX3 and/or P-47 in nine phylogenetically distinct species including *Paenibacillus larvae, Ricketsiella grylli*, *Erwinia tasmaniensis*, *Arsenophonus nasoniae, Pseudomonas putida* and *Nitrobacter winogradskyi*. The P-47 family proteins from these species show a lower level of similarity to the *C. botulinum* proteins than those from *P. dendritiformis*, but they all still contain P-47 domains. Depending on the genome annotation of their respective strains, these coding sequences are either annotated with the name of their closest P-47 family protein or as hypothetical proteins. All of these species contained multiple, heterologous P-47 family proteins (i.e. within one species there would be at least two proteins showing similarity to e.g. OrfX2 and OrfX3 respectively). This raises the possibility that they are encoded in clusters, similar to the P-47 family proteins in *C. botulinum*.

The genomic arrangement of the P-47 family proteins was investigated in nine non-BoNT encoding species. In all nine species, the P-47 family proteins were clustered together, in a similar fashion to the associated non-toxin protein encoding gene cluster in *C. botulinum.* Furthermore, in 5 of 9 strains, genes encoding putative toxin genes were encoded alongside with the P-47 family clusters.

4.1.4. Analysis of the relationship between P-47 family clusters and their co-localised putative toxins

Three different types of putative toxin were identified within the five P-47 family gene clusters. These were nematicidal proteins found in *Arsenophonus nasoniae, Halomonas* sp. TD01 and *Erwinia tasmaniensis*; a Shiga-toxin A-chain component homolog found in *Ricketsiella grylli* and a partially degraded clostridial C2 toxin component II/anthrax protective antigen homolog identified in *Paenibacillus larvae*.

This finding raises the possibility that P-47 family proteins fulfil a function that is important to the activity of different types of toxin. Examining the type of toxins encoded alongside the P-47 gene cluster in the non-BoNT encoding strains allows the generation of hypotheses of their functional role. Bacterial nematicidal proteins, which were identified in P-47 family clusters in three species, are usually toxic via ingestion - typically through their action on the midgut epithelium cells (de Maagd et al., 2003). The nematode midgut contains digestive enzymes, including peptidases, with a role in breaking down food (Terra, 1996). The hypothesis that the P-47 family proteins encoded alongside the nematicidal proteins protect the toxin from these digestive enzymes – similar to the role of the associated non-toxic proteins during BoNT intoxication, should be investigated. It should be stated that the role of the OrfX proteins in toxicity via the oral route in BoNT poisoning is unclear. However, the presence of P-47 family proteins in proximity to another oral toxin in three other species suggests that its role in toxicity via the oral route is worth further investigation. It is also interesting that the nematicidal protein in *Xenorhabdus nematophilus*, to which the *A. nasoniae* protein shows homology, is produced as part of a large protein complex (Sheets et al., 2011). This complex shows no similarity to the P-47 family proteins. There is an interesting parallel here

between BoNT and nematicidal proteins, both are co-localised with two different types of toxin complex, one of which consists of P-47 family proteins while the other does not.

The function of the P-47 family proteins in the context of the other putative toxin genes is less clear. The clostridial C2 toxin homolog is degraded, along with the P-47 family cluster, indicating that it is no longer functional in *P. larvae*. Even if the putative toxin gene were intact, the role of the P-47 family gene cluster remains unclear, as there is only a homolog of a toxin-binding region, not an activity region. However, there is a transposase immediately downstream of the C2 homolog so it is possible that component I of the C2 toxin was disrupted by this mobile genetic element. Alternatively, the P-47 family proteins may have some toxic activity that is complemented by the binding activity of the C2 toxin homolog.

In the case of *R. grylli* there is a homolog of the shiga-toxin activity subunit alongside the P-47 family encoding gene cluster. If this homolog functions in a similar fashion to the canonical shiga toxin, it would be non-functional without a cogent binding domain. One hypothesis is that the P-47 family proteins act as the binding domain in this scenario.

4.1.5. Analysis of the similarity between P-47 family clusters in different species with phylogenetic context

Although it is not feasible to perform detailed phylogenetic analysis on proteins that are so distantly related, it is possible to make cautious inferences from the similarity of the protein sequences and compare them to phylogenetic information derived from 16S rDNA sequence comparison.

There appears to be two clusters of species that show similarity in at least one P-47 family protein. One of these clusters consists of *E. tasmaniensis, A. nasoniae* and *R. grylli,* while the other cluster consists of the Paenibacillus species and *C. botulinum.* The divergence between the *Halomonas, P. putida* and *N. winogradskyi* P-47 protein sequences and any other P-47 protein sequence precludes their detailed analysis.

The phylogenetic and P-47 family sequence similarity relationships between *Paenibacillus* and *C. botulinum* are coherent, indicating that the P-47 family proteins have a similar evolutionary history to the strains that encode them. However, while *E. tasmaniensis* and *A. nasoniae* are phylogenetically related, *R. grylli* is much more divergent. The stronger similarity of the P-47 family proteins between *E. tasmaniensis, A. nasoniae* and *R. grylli* compared with the phylogenetic relationship indicates that the evolutionary history of the P-47 family proteins of these organisms does not match that of the species that encode them. It is also noteworth that *Paenibacillus* species have been shown to have anti-bacterial properties against *C. botulinum*, indicating that these two species encounter each other in the natural world (Girardin et al., 2002)

There are two other strands of evidence that point to a separate evolutionary history for the *C. botulinum*/Paenibacillus and *E. tasmaniensis/A. nasoniae/R. grylli* P-47 family clusters. Firstly, they are separated by genetic similarity and synteny. *C. botulinum* and the Paenibacillus species encode either OrfX3 or a homologue of OrfX3, while *E. tasmaniensis, A. nasoniae* and *R. grylli* share the presence of a small DNA binding protein in their P-47 family cluster. Secondly, they have distinct ecological niches; *C. botulinum* and Paenibacillus species are both soil dwelling,

anaerobic saprophytes and some species of *Paenibacillus* show antimicrobial activity against *C. botulinum*. This suggests that these two species may well share or have shared the same ecological niche, providing opportunity for horizontal gene transfer. *R. grylli*, *A. nasoniae* and *E. tasmaniensis* are all either plant of insect associated bacteria – it is possible that they also share a niche which would have enabled horizontal gene transfer of P-47 family proteins between these species.

4.1.6. Summary of analysis of the P-47 family clusters

The hypotheses presented here to explain the co-localisation of P-47 and putative toxin genes are somewhat speculative due to the lack of functional characterisation of P-47 family proteins. However, the fact that this arrangement is identified multiple times with a variety of known and potential toxins, suggests a functional relationship. Further characterisation of the P-47 family proteins, their associated putative toxin proteins and the relationship between the two will shed light on the function of P-47 and it's relevance in BoNT intoxication. The determination of the crystal structure of the P-47 family proteins would allow the deduction of the function of these proteins.

It is perhaps surprising that the P-47 family proteins are so poorly characterised, considering their wide distribution and association with known and putative toxin genes. However, all of the data on the non-BoNT encoding species in which P-47 family genes were identified comes from massively parallel sequencing experiments. These types of study provide a large amount of data, much of which will only receive a cursory analysis. These findings highlight the benefit of putting these sequences into searchable databases where it can contribute to other work. However, one downside of this kind of data (generated using 2nd generation

sequencing technologies) is that it does not result in finished genomes. The P-47 family cluster identified in *P. dendritiformis* was on a small contig with no other coding sequences. Therefore it is impossible to definitively say whether or not this gene cluster is associated with a putative toxin gene.

The results presented here show for the first time that clusters of P-47 family proteins are present in a variety of species and are frequently encoded alongside putative toxin encoding genes. This result suggests that the P-47 proteins have a functional role in toxicity. The possibility of a relationship between P-47 family proteins and toxins that exert their effect via the oral route is also worthy of investigation.

4.1.7. Prediction of potential extracellular proteins of *C. botulinum*

The number of *C. botulinum* proteins predicted to be extracellular by the software described in varied from 460 to 115. This reflects the design of the different tools to favour either sensitivity or specificity in their predictions. For example, the makers of PSORTb explicitly state that they optimised their algorithm for specificity over sensitivity, which is reflected in the fact that PSORTb predicts the second lowest number of proteins to be extracellular. The CELLO, SecretomeP and SignalP algorithms gave similar results; of the 235 predictions agreed upon by a consensus of three tools, 190 were made by these three tools. Conversely, LocateP was the tool that made the highest percentage of unique predictions (i.e. no other tool predicted that protein as extracellular). Comparison of these predictions with the *in vitro* determined proteins showed that all the tools lacked

both sensitivity and specificity. However, a full comparison of the predictions with the *in vitro* determined extracellular proteins will take place in section 4.2.5.

4.2. Proteomic investigation of *C. botulinum* to establish protein profiles associated with toxin producing strains

The whole supernatant proteome of *C. botulium* strains producing type A (ATCC 19397) and type B (NCTC 7273) toxin were investigated using bottom up proteomics coupled with LC-MS/MS. Before this work could be carried out, optimisation of protein precipitation methods had to be carried out. This involved selecting the precipitant, and subsequently, the concentration of that precipitant that gave the best representation of the supernatant proteome. Additionally the growth curve of *C. botulinum* was established to allow sampling from suitable time points. The supernatant proteome LC-MS/MS data was analysed for similarities between the two strains and for the presence of potential novel virulence factors. The relationship between protein cost and virulence was examined for extracellular proteins. Additionally, experimentally identified supernatant proteins were compared with *in silico* predictions of extracellular proteins. The generated data were used to identify specific gel sections that contained the toxin and associated non-toxic proteins (ANTPs). Corresponding gel sections were analysed from 22 additional clinical strains of *C. botulinum* to investigate the production of BoNT and the ANTPs in a range of strains.

4.2.1. C. botulinum and C. sporogenes growth curves

While the OD600 of the cultures decreased dramatically after around 12 h, the cell count of the ATCC 19397 culture only decreased marginally. The fact that the cell counting method used here does not discriminate between live and dead cells is likely to have contributed to this discrepancy. Similar findings were reported in the literature suggesting that cell density and optical density are only proportional in the positive growth stages (Monod, 1949). The number of viable cells (e.g. estimation of number of colony forming units) would be expected to correlate more closely with the optical density, although viable counts were not performed.

The decrease in OD and total cell concentration between 12h and 24 h indicates that cell lysis and sporulation may have taken place. Botulinum toxin production has long been linked to autolysis (Bonventre & Kempe, 1960; Rao et al., 2007), so it is expected that peak supernatant toxin concentration would only be reached after the culture had entered stationary phase (Bradshaw et al., 2004). However, the purported role of autolysis in the release of toxin into the supernatant means that the peak toxin and post-peak supernatant toxin concentrations will contain significant amounts of intracellular proteins. An attempt to quantiy the impact of these proteins on the constitiution of the supernatant proteomeis described in section 4.2.5.

4.2.2. Protein precipitation from *C. botulinum* and *C. sporogenes* culture supernatant

There were two problems with acetone precipitation; the resulting precipitant was incompatible with BCA assay quantification and there was a low yield when the

quantity was assessed densitometrically. The cause of these two problems with acetone precipitation is unclear. The acetone precipitation protocol resulted in a dark brown precipitant, compared with the off-white pellet of the TCA protocol. There was also a translucent, gel-like substance that was difficult to re-suspend. Other workers using acetone to precipitate proteins from bacterial culture supernatant have observed a similar phenomenon (http://life-sciences-forums.com/index.php?topic=12011.0). It is surprising that acetone precipitation results in this interference while TCA doesn't, as both acetone and TCA precipitate proteins via the same mechanism i.e. hydrophobic aggregation (Sivaraman et al., 1997). During acetone precipitation, polypropylene tubes have to be used, this was the case during these experiments, but the possibility of a contaminant being leached from the tubes would need to be controlled for in a more thorough examination of this problem. The impact of further clean up steps on acetone precipitation could also be further investigated, however, the excellent yield and breadth of proteins obtained by TCA precipitation made this unnecessary in this case.

A 5% (v/v) volume of TCA was used to precipitate proteins from the culture supernatant as it precipitated significantly more (t-test p-value < 0.05) protein from 16 h culture than 10% or 20% TCA and significantly more protein from 96 h culture than 20% TCA.

There were two problems in the comparison of the concentration of protein in the total supernatant with the amount of protein precipitated from the supernatant. Firstly, the BCA assay was instantly saturated by the whole supernatant, even with serial dilution, meaning that this assay could not be used for the quantification of

the whole supernatant. Secondly, the Bradford assay overestimated the amount of protein in the whole supernatant by a factor of 3-fold.

The BCA assay quantifies polypeptides of more than 3 amino acids in length (http://www.piercenet.com/browse.cfm?fldID=876562B0-5056-8A76-4E0C-B764EAB3A339). When the BCA assay is used to quantify total supernatant protein, the high level of proteinaceous content in the TPGY (55 mg/ml) significantly exceeded the amount of bacterial protein and saturated the assay. This saturation does not occur in the TCA precipitated samples as only proteins that have a molecular weight greater than 10 kDa are precipitated by TCA (personal communication, Thermo Scientific technical support). The majority of proteinaceous material in TPGY is below this molecular weight and is therefore not precipitated and subsequently quantified when a sample has been TCA precipitated. This explains why, when TCA is used to precipitate the protein from sterile TPGY, only around 0.5 mg/ml of protein is obtained despite having 55 mg/ml of proteinaceous content. This could be empirically confirmed if the proteins and polypeptides in TPGY were examined on a 10% Bis-Tris gel with MES running buffer, conditions that provide good resolution at low molecular weights.

When total supernatant protein is quantified using the Bradford assay a concentration is obtained, i.e. the assay is not saturated as with the BCA assay. This is because the Bradford assay only quantifies proteins that are greater than 3 kDa (http://www.piercenet.com/browse.cfm?fldID=876562B0-5056-8A76-4E0C-B764EAB3A339). The amount of protein in TPGY that is greater than 3 kDa results in an overestimation compared with the TCA-precipitated protein sample but not in a saturation of the assay.

These results highlight the high protein content of TPGY, and shows that *C. sporogenes* culture is only metabolising a fraction of the total protein in the medium.

Based on these conclusions it might be expected that when the BCA assay is used to quantify a total supernatant sample that has been processed with a Molecular Weight Cut-Off filter to reduce the concentration of proteins <3 kDa, a similar result to the Bradford assay of the total supernatant would be obtained. However, this was not the case and the BCA assay was still saturated by TPGY polypeptides. This is likely because if there is a large amount of protein that should pass through the MWCO filter, the filter becomes clogged and it's efficiency is reduced. Using multiple MWCO filters per sample may improve this efficiency but is not cost-effective.

Due to the technical issues detailed above, it is difficult to draw firm conclusions on what proportion of bacterial protein is being precipitated by the addition of TCA. Previous workers have reported that it is necessary to precipitate supernatant proteins prior to quantification due to the high salt content (Schwarz et al., 2007). However, the similarity of the protein profile of the total supernatant and the precipitated protein indicates that precipitated proteins are an accurate representation of the supernatant proteome. This is unsurprising as the majority of bacterial proteins have a molecular weight >10 kDa.

4.2.3. Determination of toxin concentration in the culture supernatant using endopeptidase assay

As the focus of this study is the toxin and toxin complex, an endopeptidase-assay that quantifies botulinum neurotoxin activity was employed (Jones et al., 2008) to identify the time point at which supernatant toxin concentration was highest. Previous tests have shown that proteases formed by ATCC 19397 did not cause a false positive reaction in this type A endopeptidase assay (Sharma, 1999).

At 0 h, residual toxin activity was detected in the culture supernatant, this residual activity then increased 16 fold to peak at 24 h. The active toxin at 0 h was transferred to the culture in the 1 ml inoculum.

The endopeptidase assay showed that supernatant toxin concentration peaked at 24 h and then decreased steadily until 96 h when the experiment was terminated. This is in direct contrast to previous reports by Bradshaw et al. (2004) that have shown that, while supernatant toxin concentration varies depending on strain and culture conditions, in all tested conditions supernatant toxin concentration and activity (as measured by mouse bioassay) was equivalent or higher at 96 h compared with 24 h (Bradshaw et al., 2004). One potential explanation for the decrease in toxin observed between 24 h and 96 h in this work is that the supernatant toxin is being degraded by extracellular proteases. Characterisation of the supernatant proteome would help to identify enzymes that could be responsible for this. However, there are no reports of *C. botulinum* producing a protease that can degrade the toxin while it is in the toxin complex. One possibility is that the pH of these cultures increased to an unusually alkaline level, resulting in toxin complex disassociation and subsequent degradation of the toxin complex. A

definitive answer to this discrepant result is not available from the current data and further work to confirm this phenomenon and investigate whether the culture is becoming alkaline are required.

4.2.4. Analysis of *C. botulinum* supernatant proteome

The genome sequence of *C. botulinum* A ATCC 19397 was used as a reference in the analysis of its proteome. However, the *C. botulinum* B strains that had finished genomes at the time of this study were not available to us. Therefore, the genome of another BoNT/B encoding proteolytic strain, *C. botulinum* B Okra was used as the reference for the identification of *C. botulinum* B NCTC 7273 supernatant proteins. This genome was used after intial work showed that it provided more matches than any other *C. botulinum* genome then available. NCTC 7273 was chosen for analysis as it is the UK type strain for *C. botulinum* B and is used by the Botulinum Reference Lab in the development of new assays.

There were 80 additional proteins identified in the supernatant of *C. botulinum* A ATCC 19397 than in the supernatant of *C. botulinum* B NCTC 7273. This is likely because the whole genome sequence of ATCC 19397 was available while another *C. botulinum* B genome, Okra, was used as the reference for the analysis of the NCTC 7273 supernatant proteome. When proteomic data is analysed, the peptides identified by the mass spectrometer are compared against a database of *in silico* peptides from the reference genome. If there are protein sequence differences between the analysed strain and the reference genome then the experimentally identified peptides will not match the peptides in the *in silico* database and the protein may not be identified (depending on how divergent the reference and experimentally identified protein are). The NCTC 7273 proteome

was also analysed using ATCC 19397 as a reference, this returned fewer protein matches than the Okra genome. Previous work has found that the investigation of the proteome of un-sequenced bacteria and fungi results in fewer identifications than when analysing sequenced strains (Shabbiri et al., 2013; Cobos et al., 2010). This emphasises the importance of the availability of appropriate reference sequences for bottom up proteomic analysis.

More proteins may have been identified in the NCTC 7273 supernatant if a comparison had been performed against a database of proteins from all the sequenced *C. botulinum* strains. The protein sequences from every available genome sequence could have been clustered, e.g. using CD-HIT (Li & Godzik, 2006) to cluster at 85% similarity, a single representative protein taken from each of these clusters, and the NCTC 7273 proteome could be compared against a wider range of *C. botulinum* proteins, likely resulting in more matches. However, this approach requires a level of bioinformatic sophistication that was not available at the time of analysis. A simpler database that contains all the proteins from all the sequenced strains would contain multiple versions of highly similar proteins. This can lead to a decrease in sensitivity as the lack of unique peptides prevents unambiguous identification.

The supernatant proteome of *C. botulinum* A1 ATCC 19397 was investigated at 24 h and 96 h while the supernatant proteome of *C. botulinum* B NCTC 7273 was investigated at 24 h. The supernatant proteome of *C. botulinum* A1 ATCC 19397 was quite stable between 24 h and 96 h with 65% of identified proteins being detected at both time points. Of the 72 proteins that were only identified at one time point, 63 were identified at 24 h only. This is to be expected as there would

be less protein synthesis between 24 h and 96 h than between 0 h and 24 h. Notably, pathogenesis proteins were the largest group in which all members were identified at both 24 h and 96 h. There were no functional groups that showed a strong pattern of presence at only one time point. This indicates that the 96 h proteome is largely similar to the 24 h proteome but with degradation occurring across the proteome reducing the number of identifiable proteins at 96 h. Protein fate, which includes proteases, was the third largest group at both 24 h and 96 h highlighting the highly proteolytic nature of *C. botulinum*. One of these proteases may be responsible for the decrease of BoNT detected by the endopeptidase assay.

Previous investigation of the supernatant proteome of *C. botulinum* identified 10 proteins in addition to the botulinum toxin and toxin complex (Cheng et al., 2008). The work presented here confirms the presence of some of the previously identified proteins (including Clp protease, nlpc/p60 family protein, ornithine carbamoyltransferase) and, in addition, builds on them, showing that *C. botulinum* has a much more complex and diverse supernatant proteome than previously thought. While the sophisticated mass spectrometry method used here would explain why more proteins were identified from the supernatant in this study than in previous work (Cheng et al., 2008) it does not explain why fewer protein bands were present by SDS-PAGE. This could be due to the use of a more minimal media, Toxin Production Media, resulting in the production of fewer extracellular enzymes. It has been hypothesized that cell density is an important factor in BoNT production (Zheng et al., 2013; Schantz & Johnson 1992), it could be that differences in media composition effect the cell density which effects expression of supernatant proteins.

In addition to the well-characterised botulinum toxin and neurotoxin associated proteins, which were previously reported by Cheng et al. (2008), the putative toxin activating enzyme, clostripain, was detected in this study. Proteins exhibiting similarity to pathogenicity factors from other organisms were also identified by comparison with MvirDB. This database was chosen as it is a comprehensive collection that includes proteins from other, more specific databases (e.g. Tox-Prot and the Virulence Factor Database). One potential novel virulence factor was identified in both *C. botulinum* A ATCC 19397 and B NCTC 7273. This was a thermolysin metallopeptidase that showed 87% amino acid identity across the entire length of its sequence between the ATCC 19397 and NCTC 7273 proteins. These proteins showed similarity to the *C. perfringens* lambda toxin that has been shown to activate the *C. perfringens* epsilon proto-toxin by cleaving 2 kDa from the N-terminus (Gibert et al., 2000; Minami et al., 1997). It is only identified in strains of *C. perfringens* that also encode the epsilon toxin, indicating that it is functionally related to the epsilon toxin (Matsushita & Okabe, 2001). It is possible that the *C. botulinum* thermolysin metallopeptidase plays a similar role in activating the botulinum neurotoxin by cleaving the light chain and heavy chain. Another *C. botulinum* A ATCC 19397 extracellular protein with similarity to a previously identified virulence protein was the ATP-dependent Clp-protease that showed similarity to a Clp-protease from *Listeria monocytogenes*. The Clp-protease is involved in phagosome escape and activation of listeriolysin-O (Gaillot et al., 2000). In *C. botulinum* it could be involved in modulating the host response in e.g. wound botulism. Host immune system cells recruited to respond to the wound infection could be disrupted by bacterial effectors, this type of mechanism is frequently used by bacteria and can result in disease or chronic infections (Finlay & McFadden, 1996). The presence in the *C. botulinum* supernatant proteome of an analogue to a vital pathogenicity factor from another organism highlights the potential

involvement of hitherto uncharacterised proteins in the virulence of *C. botulinum.* Similarly, a *C. botulinum* B NCTC 7273 protein that showed similarity to a characterised virulence factor was a collagenase protein. These proteins may play an important role in wound botulism due it their ability to destroy tissue integrity in an infected host (Harrington, 1996).

There are also a large number of degradative enzymes including proteases and lipases. There were 19 proteins involved in protein fate in the core supernatant proteome of ATCC 19397 and NCTC 7273, of which 11 were involved in protein degradation. The high number of proteases identified in the supernatant reflects the large number of protease encoding genes identified in the genome of *C. botulinum* A Hall strain (Sebaihia et al., 2007). The identification of these proteins as expressed, allows a much more vivid picture of their role in the ecology of the organism, than their presence in the genome alone. The expression of the protein fate proteins was highly conserved, with 79% of the protein fate proteins identified in the supernatant of either organism being present in the supernatant of both organisms. The conserved expression of the protein fate proteins indicates that they are important to the metabolism of the organism. Of the three functional groups that are more highly conserved than protein fate two are very small and the other is pathogenesis. This conservation makes it less likely that one of these proteins is responsible for the decrease in BoNT seen in the endopeptidase assay. If the toxin was being degraded by a protease with highly conserved expression the decrease would have been reported in previous work (Bradshaw et al., 2004). In addition, if there were a protease responsible for degradation of BoNT it would also have to be capable of degrading the complexed form of the toxin which would be present at the pH of a post-exponential *C. botulinum* culture. However, it is deemed highly unlikely that the organism would produce such an enzyme.

Cheng et al. (2008) compared the lethal dose of the toxin (in both complexed and uncomplexed form) with that of their 'crude' extract in an intra-gastric mouse model. The crude extract had at least 10 additional proteins (as identified by mass spectrometry). They found no significant difference in lethality between the crude extract and the toxin complex. The exact protocol for obtaining the 'crude' extract is unclear from their publication; however, it appears to be an acid precipitation from a 96 h culture, with no further purification steps. This is surprising as the crude toxin sample they analysed by 1D-GE is much less complex than the one analysed in this work, with approximately 17 protein bands identified compared with around 40 identified in this work. It could be that the Toxin Production Medium used in the experiments of Cheng et al. (2008), results in less supernatant proteins than the TPGY media used here. Toxin Production Medium is less nutritious than TPGY, with 2% casein hydrolysate and 1% yeast extract compared to 5% trypticase, 0.5% bacto peptone and 2% yeast extract in TPGY. The effect of medium composition on toxin production has been widely investigated, with amino acids and peptides found to be most important for BoNT regulation (Schantz & Johnson, 1992). However, the effect of medium composition on the diversity of the broader proteome has not been investigated. In the light of these findings, it would be interesting to explore the link between medium composition, toxin production and the diversity of *C. botulinum* supernatant proteins. One hypothesis to investigate is that, in more amino acid limited culture conditions (such as Toxin Production Medium), *C. botulinum* produces more BoNT at expense of the various degradative enzymes identified here.

There were a large number of proteins in the supernatant of *C. botulinum* that would typically be thought to be intracellular proteins, e.g. proteins involved in

electron transport, protein synthesis or transcription. This is similar to other studies investigating bacterial supernatant proteomes (Kaakoush et al., 2010; Pocsfalvi et al., 2008; Walz et al., 2007). These were detected due to cell autolysis during growth that results in intracellular proteins being released into the extracellular milieu.

4.2.5. Comparison of predicted and experimental supernatant proteome

As mentioned above, not all the proteins identified in the culture supernatant are expected to be actively secreted by the organism, some will be there as a result of cellular autolysis. Therefore, *in silico* tools were used to predict which supernatant proteins had motifs typical of secreted proteins. An analysis of which tool provides the best prediction of the supernatant proteome was also carried out. The 5 tools used predicted between 10 and 27 of the 207 *C. botulinum* A ATCC 19397 supernatant proteins (4.8-13.0%) to be extracellular. The number of correct predictions was positively correlated with the total number of predicted proteins. It is difficult to assess how successful individual tools have been due to the following uncertainty; are the 87-95% of proteins in the supernatant that are not predicted to be extracellular just there as side effect of lysis, or are they secreted but just not predicted as such by tools? This dataset is not designed to answer this question, but other, previously published datasets can address this question more directly. Bumann et al., (2002) investigated the supernatant proteome of *Helicobacter pylori* using methods explicitly designed to minimise autolysis. They identified 26 secreted proteins using mass spectrometry. When the genome of *H. pylori* was analysed using PSORTb, 241 proteins were predicted to be extracellular. When the predicted proteins were compared with the empirically identified extracellular

proteins, only 3 of the experimentally identified proteins were predicted to be extracellular. This means that even in a best-case scenario for reducing intracellular contamination, PSORTb predicts only around 10% of supernatant proteins to be extracellular.

As a general rule, sensitivity (number of extracellular proteins correctly predicted) was only improved with a loss of specificity (number of predictions identified experimentally). The specificity was generally low, with between 4.9% and 15.7% of predicted proteins being identified and the mean percentage across the 5 tools being 7.8%. The sensitivity was also low, with a range of 4.8% to 13.1%, and an average of 9.7%. The most sensitive tool was CELLO, which correctly predicted 13.1% of experimentally identified extracellular proteins. The least sensitive tool was PSORTb, which only correctly predicted 4.8% of the experimentally identified extracellular proteins.

The most specific tool was LocateP, which runs each protein sequence through a series of Hidden Markov Models each of which predicts whether the protein has a different sub-cellular location. This approach is not significantly different to the other sub-cellular location prediction tools, which generally employ a machine-learning algorithm (such as Hidden Markov Models or Support Vector Machines) trained on a dataset of proteins of known sub-cellular location. However, the predictions of LocateP were notably better than any other tool, with 15.7% of it's predictions being identified in the culture supernatant while the next best accuracy was 6.4% (PSORTb). While LocateP had the highest specificity, it had the second worst sensitivity, with only 8.7% of extracellular proteins correctly predicted. CELLO had the highest sensitivity (13%), but any advantage this might provide is

lessened by its having the second worst specificity, with only 5.9% of it's predictions being found extracellularly.

In future work, if an *in silico* tool was sought to predict supernatant proteins an assessment of whether sensitivity or specificity was more desirable would need to be carried out. Even then, the low success rate of even the best of these tools would need to be taken into consideration when interpreting the results of such an analysis.

The CELLO, SecretomeP and SignalP algorithms gave similar results; of the 235 predictions agreed upon by a consensus of three tools, 190 were made by these three tools. Conversely, LocateP was the tool that made the highest percentage of unique predictions (i.e. no other tool predicted that protein as extracellular). Of the 18 proteins that LocateP correctly predicted as being extracellular, 5 were not predicted as extracellular by any other tool. Two of these proteins were transport proteins, two were cell envelope proteins and one was hypothetical. The two transport proteins could be expected to be extracellular as substrate binding proteins are often released into the extracellular milieu in order to scavenge resources. Cell envelope proteins are obviously required to pass through the plasma membrane in order to fulfil their role on the outer surface of the bacterium, and so it is expected that they would have motifs typical of extracellular proteins that are responsible for their trans-membrane transport (Van Wely et al., 2001).

Although LocateP had the highest accuracy generally, it was the only tool not to successfully predict that BoNT and NTNH would be extracellular. CELLO was the only tool to successfully predict HA33, HA17 and HA70 as supernatant proteins. Of the 11 proteins predicted as extracellular by all 5 tools, 6 of them were

thermolysin metallopeptidases. These proteases are well characterised as being extracellular and would be in the training set of all these tools.

In order to investigate whether the poor performance of extracellular prediction tools was specific to *C. botulinum*, a meta-analysis of previous work characterising bacterial supernatant proteomes was carried out. PSORTb was then used to predict the supernatant proteome of organisms with experimentally characterised supernatant proteomes and the prediction results compared to the experimental evidence. PSORTb was employed as it is a widely used tool that accepts protein sequences in FASTA format while other tools such as LocateP only work for pre-computed genomes. Four studies were identified; looking at the supernatant proteomes of *Listeria monocytogenes* (Dumas et al., 2009), *Campylobacter concisus* (Kaakoush et al., 2011), methicillin resistant *Staphylococcus aureus* (Burlack et al., 2007) and *Helicobacter pylori* (Bumann et al., 2002). The *L. monocytogenes* and MRSA studies were performed using a 2DGE-MALDI approach while the *Campylobacter and Helicobacter* studies were performed using a 1DGE-LC-MS/MS approach. When the experimental results were compared with the predictions, 16.3% of the *L. monocytogenes*, 0.5% of the *C. concisus*, 14.9% of the MRSA and 1.2% of the *H. pylori* predictions were experimentally identified. When this is compared with the 6.4% specificity of *C. botulinum* PSORTb predictions the *C. botulinum* specificity is the median value of these five comparisons while the other results are bimodally distributed. While there is no relationship between higher specificity and technique used, the organisms with higher prediction accuracy are Gram-positive. PSORTb performs different analyses based on Gram-stain predictions, from this small sample, it appears that the Gram-positive algorithms perform significantly better. The approach taken to develop the Gram-positive and Gram-negative PSORTb algorithms was identical,

with the only difference being the datasets used to train the Support Vector Machines. The results presented here indicate that the training dataset for the Gram positive organisms was better than that for the Gram negative bacteria. This highlights one of the key limitations of machine learning approaches for sub-cellular localisation; the algorithm can only be as good as the data used to train it. The fact that experimentally identified extracellular proteins in the literature were not identified as extracellular by PSORTb indicates that these proteins were not in it's training dataset (PSORTb v3 has 97.3% specificity and 92% sensitivity at predicting the subcellular location of proteins in it's training dataset). Expansion of the training dataset used to develop these algorithms is likely to improve the results obtained here.

It is difficult to accurately predict the sub-cellular location of bacterial proteins due to the varied nature of bacterial protein export systems (Holland, 2010). While some systems, such as the Sec pathway, are well understood (Chatzi et al., 2013) our understanding of most of these systems is still in a state of flux. It is difficult to implement *in silico* methods to predict the targets of poorly understood systems. Therefore the majority of tools rely on a machine learning approach that can only be as good as the training data set used in it's development. Until more experimental evidence is accrued, both through bottom up proteomic approaches like those evidenced here and targeted investigation of secretion mechanisms, secreted protein prediction is unlikely to significantly improve.

4.2.6. Relationship between extracellular protein cost and association with virulence

This section addresses the hypothesis that extracellular proteins are metabolically cheaper than non-extracellular proteins and whether extracellular protein cost is an indicator of importance to ecological niche/pathogenesis.

Previous work has presented evidence that there is an evolutionary pressure on bacteria to reduce the metabolic cost of extracellular proteins (Smith & Chapman, 2010). However, this work was performed using sub-cellular locations predicted by PSORTb, which, as the previous section showed, is not an accurate method for predicting extracellular proteins. Here, the average metabolic cost of experimentally confirmed supernatant proteins was compared against the average metabolic cost of non-extracellular proteins. A small but significant difference was identified between the extracellular and non-extracellular proteins. This difference is the first experimental evidence that suggests extracellular proteins are metabolically less expensive than intracellular proteins. The difference in cost between the extracellular and non-extracellular proteins is smaller than was found by Smith and Chapman, 1.2 HEPBs (High Energy Phosphate Bonds) per amino acid compared with 2.7 HEPBs. However, there were only 16 predicted extracellular proteins in the *E. coli* K-12 genome they used in their work and the small sample increases the chance of skewed results (Tversky & Kahneman, 1971).

A major caveat to this analysis is that the proteins that are identified here as extracellular may be present in the culture supernatant as a result of autolysis rather than deliberate secretion.

Despite being generally cheaper, some extracellular proteins were large and metabolically expensive. These include BoNT that is the second most expensive protein in the culture supernatant in terms of total cost and the 14th most expensive protein in terms of average amino acid cost. All the other BoNT complex proteins and clostripain are within the top 15% of either the total or average amino acid cost. These are the six *C. botulinum* supernatant proteins that are annotated as having a role in pathogenesis.

This pattern of virulence proteins being among the most expensive proteins in the supernatant was continued in other organisms for which data was available. In two studies on *Staphylococcus aureus*, 8 of 8 and 7 of 11 extracellular pathogenicity proteins were in the top 25% most expensive extracellular proteins (Burlack et al., 2007; Pocsfalvi et al., 2008). These proteins included extracellular enterotoxin L, entertoxin H, enterotoxin B and various hemolysins. Analysis of data from studies on *Listeria monocytogenes* and *Campylobacter concisus* found that 3 of 4 and 3 of 6 extracellular pathogenicity proteins were in the top 25% expensive proteins respectively (Dumas et al., 2009; Kaakoush et al., 2010). This suggests that extracellular virulence proteins, including toxins, are more expensive than the average extracellular protein for various bacterial pathogens.

Smith and Chapman hypothesised that it was evolutionary pressure on the organism that resulted in extracellular proteins being metabolically cheaper. The fact that these expensive proteins (BoNT and the ANTPs) are vital to the ecological niche of the organism could support this hypothesis because there is an opposing evolutionary pressure on the organism to conserve these proteins. This conservation of amino acid sequence will prevent the reduction in cost of these

important extracellular proteins while proteins that play a less pivotal role for the organism become cheaper under the evolutionary pressure postulated by Smith and Chapman. One example of cheaper extracellular proteins in *C. botulinum* is a copper chaperone (A7FTI3) that is in the bottom 25% of extracellular proteins in terms of total and average per amino acid cost.

It is also interesting that the 3 of the 4 OrfX proteins are in the top 15% of proteins in terms of total or average per amino acid cost. There has been limited phenotypic characterisation of these proteins, but their expense relative to the other *C. botulinum* supernatant proteins indicates that they are being conserved.

Protein cost could be used as a guide for investigating the supernatant proteome of uncharacterised pathogens and other organisms. Proteins that are important to the ecological niche of the organism are likely to be more conserved as there is a strong pressure on the organism to maintain the function of the important protein. This conservation will counteract the evolutionary pressure to reduce protein cost resulting in more important proteins tending to be more metabolically expensive.

It would be preferable to be able to compare the cost of the supernatant proteins with the cost of the proteins identified in whole cell proteomic work rather than the non-extracellular proteins used here. However, this dataset was unavailable in this work.

4.2.7. Identification of toxin complex proteins in clinical strains of *C. botulinum* by LC-MS/MS

The expression of BoNT and ANTPs was investigated in clinical strains of *C. botulinum*. The vast majority of proteomic work investigating *C. botulinum* has focused on a small number of type strains. Here, we utilise the unique clinical strain collection of Public Health England to improve understanding of the role of the botulinum toxin complex in clinical strains associated with food, wound and infant botulism.

When the supernatant proteome of ATCC 19397 was analysed by 1D-GE/LC-MS/MS, no BoNT peptides were identified in the 115-200 kDa gel section, while peptides matching the BoNT heavy chain were identified in the 97-115 kDa gel section and peptides matching the light chain were identified in the 50-55 kDa gel section. This shows that all the BoNT in these culture supernatants was in the 'nicked' form, with the light chain and heavy chain only joined by a disulphide bond that would be broken by the reducing conditions present in the buffer. The fact that all the toxin was present in the nicked form indicates highly active toxin activating enzymes in the culture supernatant, of which clostripain has been previously implicated as one (Dekleva & Dasgupta, 1990). HA70 also showed post-translational cleavage, as has been previously characterised (Inoue et al., 1996), into two subunits of 50 kDa and 20 kDa. This post-translational nicking of HA70 has recently been implicated in the formation of the hetero-dodecameric HA sub-subcomplex (Lee et al., 2013). This nicking suggests that the toxin complex present in the supernatant of ATCC 19397 is of the form suggested by Lee et al (2013). NTNH, HA33 and HA17 appeared un-cleaved. This result contrasts with previous reports that NTNH/A is nicked to give fragments of 13 kDa and 106 kDa

indicating that the previously reported cleavage is not a universal phenomenon (Fujita et al., 1995)

When the supernatant proteome of NCTC 2012 was analysed, BoNT, P-47, OrfX1 and OrfX3 were not identified. NCTC 2012 was isolated from the first reported outbreak of botulism in the UK (food botulism associated with duck paste in Loch Maree in 1922) and has been shown to produce 100-1000 fold less toxin than BoNT/A1 producing strains (Tepp et al., 2012). The reason for this reduction in toxin production is unclear, although yield has been shown to be increased 10-100 fold by culturing in modified Mueller-Miller medium (Tepp et al., 2012). The fact that OrfX2 and NTNH were identified in the culture supernatant indicates that it is unlikely that NCTC 2012 lost the plasmid that encodes the toxin gene cluster. Two component systems are thought to be involved in regulation of the BoNT/A1 gene (Connan et al., 2012). Variation between A1 and A3 producing strains in the two component systems involved in toxin regulation could explain the difference in production. It is noteworthy that changes in the media composition, which would be detected by two component systems (Connan et al., 2012), result in higher levels of toxin expression in A3. This indirectly supports the hypothesis that two component systems are involved in toxin regulation. A more detailed picture of the transcriptional activity of NCTC 2012 under various growth conditions, along with characterisation of the two component systems involved in BoNT/A3 gene regulation would assist in understanding the lower toxin production in this strain. As BoNT/A3, P-47, OrfX1 and OrfX3 were not identified in the supernatant of the type strain NCTC 2012, gel fragments corresponding to their predicted molecular weights were analysed in clinical strains encoding OrfX.

This work is the first to examine the production of the botulinum toxin and associated proteins in a variety of clinical isolates. The method used identified toxin in 20 of 22 clinical isolates. In all 20 of these isolates, the type of toxin identified matched the type of toxin identified by PCR. No unique peptides matching the toxin were identified in one type A strain from 2004 and one type B strain from 2009, both strains were associated with wound botulism. Interestingly, these were also the only two strains for which NTNH was not detected. The presence of the toxin genes in these strains had been confirmed prior to these experiments so it is unlikely the toxin gene has been lost. Also, the presence in the culture supernatant of other neurotoxin-associated proteins indicates that at least part of the toxin gene cluster is operational. As BoNT and NTNH are encoded as part of the same operon, one potential hypothesis is that either the RNA polymerase binding site on the DNA, or the ribosome-binding site on the mRNA is disrupted in these strains. This would disrupt the *bont-ntnH* operon while leaving the *ha* operon functional. It is also possible that these strains showed a different dynamic of toxin production through the time-course than the reference strain, with toxin being present at a detectable concentration at times other than the 24 h point sampled here. The proteomic analysis of the toxin complex components produced by a BoNT/G producing organism identified BoNT, NTNH, HA17 and HA70 (Terilli et al., 2011). This indicates that expression of only some parts of the toxin complex could be a biological phenomenon. In the case of BoNT/G, the lack of HA33, hypothesised to be essential in transport of the toxin across the gut epithelium (Sugawara et al., 2010; Lee et al., 2013) could explain why this toxin type is not associated with food borne botulism. In contrast, the HA33 protein was ever-present in the clinical *C. botulinum* strains investigated here. The work of Lee et al. (2013) also raises the question of whether strains that don't produce all the HA components of the toxin complex would form a functional HA sub-complex.

However, further technical validation of this result in the *C. botulinum* strains investigated here would be the first step to investigating these exceptions. Three bivalent strains that encoded toxin types A and B were also investigated but only a single toxin type was identified in each strain. This follows the usual pattern for bivalent strains where either all or the majority of toxin produced is of one type (Peck, 2009).

As mentioned above, prior to proteomic analysis, PCR amplification a fragment from two genes for each toxin complex was used to determine whether a strains encoded HA or OrfX proteins; 19 of the clinical isolates investigated only encoded the HA complex with the other 3 clinical isolates encoding both HA and OrfX encoding genes.

Of the 19 HA only encoding clinical isolates, all three HA proteins were identified in 13 strains, with HA70 missing from two and HA17 missing from four isolates. Before investigating alternative hypotheses for the absence of these proteins further technical validation of these results should be carried out. Despite these exceptions, production of the HA toxin complex was highly conserved among clinical strains of *C. botulinum* associated with food, wound and infant botulism. Lee et al. (2013) found that the entire HA sub-complex is required to facilitate absorption of the toxin through the gut epithelium. Therefore, it is deemed likely that the absence of HA70 and HA17 from two and four isolates respectively is a technical issue rather than a biological one. This is because; according to the model of Lee et al. (2013) the absence of these proteins would render the toxin incapable of passing through the gut epithelium and two of the strains that lacked HA17 were associated with food-borne botulism.

Three bivalent HA/OrfX strains were investigated, none of which were found to produce proteins from both toxin complex types with two strains producing HA proteins and one strain producing OrfX proteins. One strain that only encoded OrfX genes was also investigated. Of the two isolates that produced OrfX proteins, both were associated with infant botulism. Of the four OrfX cluster proteins (P-47, OrfX1, 2 and 3), OrfX2 was identified in the culture supernatant of both strains while OrfX3 was only identified in one. Neither OrfX1 nor P-47 was identified in either culture supernatant. This contrasts with previous work that has shown that in a *C. botulinum* A2 (i.e. OrfX) encoding strain, OrfX2, OrfX3 and P-47 were identified while OrfX1 was not (Lin et al., 2010). Although only a small number of OrfX producing strains were analysed here, these results indicate that OrfX proteins are produced by clinical strains of *C. botulinum.* However, not all the OrfX cluster proteins were produced to levels detectable by the methodology used here. It should also be mentioned that there could be cleavage of the OrfX cluster proteins that would result in their being a different molecular weight and thus not migrating to the section of the SDS-PAGE gel selected for investigation. Further work to resolve this would involve a full supernatant proteome characterisation of these OrfX encoding strains.

Another interesting facet of the results presented here is the insight into the toxin production of strains that encode two toxin types. There were three strains investigated that encode two complete *bont* genes (A5(B) strains encode an incomplete *bont/B*). NCTC 2916 is known to only produce BoNT/A (Bradshaw et al., 2004) and the results presented show that it produces the HA toxin complex proteins. This result is particularly interesting because in NCTC 2916 the *bont/A1* is encoded as part of an OrfX gene cluster and the HA proteins are encoded as

part of the same cluster as the silent *bont/B* gene (Rodriguez Jovita et al., 1998). This indicates that there is coordinated expression from two genomically separate toxin complex clusters, one that produces BoNT/A and NTNH while the other produces the HA proteins. The mechanism by which this expression is regulated would be very interesting to probe. There are two other strains that were shown to encode two toxin genes by PCR, a food-botulism strain (H063740588) and a wound botulism strain (H091640054) that both encoded toxin types A and B. However, proteomic analysis of the supernatant of these strains showed that only one toxin type was produced to detectable levels, BoNT/A in both strains. The fact that they seem to only produce one toxin type, rather than producing a biologically inactive secondary toxin type agrees with and extends previous findings (Singh et al., 2013; Rodriguez Jovita et al., 1998; Hutson et al., 1996).

In summary, toxin complex proteins are expressed in all clinical isolates of *C. botulinum* investigated here. In the majority of strains, all toxin complex proteins were detected. Expression of OrfX proteins appear to be not as well maintained as that of HA proteins, although the small number of OrfX encoding strains analysed here prevents firm conclusions from being drawn.

4.3. Genomic diversity and toxin complex type of clinical isolates causing botulism in the UK

fAFLP was employed to investigate the genomic diversity of 38 isolates of proteolytic *C. botulinum* and neurotoxigenc *C. butyricum* from clinical cases of botulism in the UK.

The resulting clusters, representing the genetic diversity of BoNT producing organisms, were compared with the toxin complex type (i.e. HA and/or OrfX) of each strain and the type of botulism (i.e. food, wound or infant) that each strain was associated with.

The AFLP results obtained here are consistent with results from earlier work differentiating *C. botulinum* strains from different countries using AFLP, MLVA and MLST (Hill et al., 2007; Macdonald et al., 2008; Jacobsen et al., 2008; Hill & Smith, 2013). Similarly to previous work, they show the genomic diversity present in *C. botulinum* (Carter et al., 2009). Specific similarities include distinct clustering of bivalent, AB strains; the distant relationship between Ba4 657/NCTC 2012 strains and the other proteolytic *C. botulinum* and the very distant relationship between the proteolytic *C. botulinum* and the toxigenic *C. butyricum*.

There was concordance between genomic similarity and toxin complex profile. Of the 25 strains in clusters 1, 2, and 3 there were 23 that encoded HA only, while two were bivalent, encoding both HA and OrfX gene clusters. The two bivalent strains were closely related to each other and distantly related to the other cluster 2 strains. Cluster 4 consisted of 5 strains, of which 4 encoded OrfX only while 1 was bivalent, with both HA and OrfX encoding genes. This separation of strains

encoding HA and OrfX strains is seen in previous work, although toxin subtype has to stand in as surrogate for toxin complex type (i.e. A2, A3, A4 are OrfX encoding; A1 and B are HA encoding) (Hill et al., 2007; Fillo et al., 2011; Luquez et al., 2012).

There is also correlation between the type of botulism caused and genomic similarity. Of the 20 clinical strains in clusters 1 and 2 there were 18 wound botulism strains and 2 infant botulism strains. The infant botulism strains in clusters 1 and 2 were distantly related to the other strains in their respective clusters. There was only 1 clinical strain in cluster 3, an isolate associated with a case of food botulism. All 5 strains in cluster 4 were infant botulism associated.

Cluster 5 consisted of 5 *C. butyricum* strains, 4 of which were toxigenic. The non-toxigenic *C. butyricum* was distantly related to the toxigenic strains that were indistinguishable from each other. Three of the toxigenic strains were from a single incident of infant botulism, while the fourth toxigenic strain was linked to a separate case of infant botulism.

Four hypotheses could explain these results.

4.3.1. Hypothesis 1 – toxin complex type directly contributes to disease type

One hypothesis is that there is a functional relationship between toxin complex type and type of botulism caused. The data presented here show that HA encoding strains can cause wound, food or infant botulism while strains that encode only OrfX are, with one exception, associated with infant botulism. The

one exceptional OrfX only encoding strain linked with food botulism was *C. botulinum* A3 NCTC 2012 which caused 8 deaths at Loch Maree in 1922. It is also interesting that while this strain (and toxin type) is known to produce low amounts of toxin (Tepp et al., 2012), it resulted in more deaths than any other incident of food borne botulism in the UK. However, this is likely due to lack of knowledge of the disease and the unavailability of the anti-toxin (McLaughlin et al., 2006). It is noteworthy that this pattern extends to the expression of toxin complex by strains that encode two toxin complex types. There was one clinical bivalent HA/OrfX strain that was associated with food-borne botulism that proteomic work showed only produces proteins from the HA cluster. There is also one bivalent HA/OrfX strain that was associated with infant botulism and proteomic analysis showed this strain to produce only OrfX proteins.

One key question in addressing this hypothesis is whether the OrfX proteins actually provide an advantage during infant botulism toxico-infection, or whether OrfX strains are just less able to cause wound and food botulism. The fact that, of the 10 infant botulism associated strains investigated here 8 encoded OrfX, is suggestive that the OrfX cluster may provide an advantage in infant botulism. One hypothesis to explain this is that the infant gut epithelium has histological characteristics that make it susceptible to disruption by a P-47 family protein complex. This kind of age-dependent susceptibility of the gut to a pathogen has been seen in rotavirus, which affects infants less than 6 years, while adults do not develop symptomatic disease (Pott et al., 2012). The changing susceptibility was due to the increased expression of the innate immune receptor for viral dsRNA in older mice (Pott et al., 2012). When looking for OrfX targets, cell receptors that are over-expressed in infants are worthy of close investigation. There is also a link between infant botulism and weaning, the infant gut epithelium shows accelerated

epithelial growth during weaning, with increases in the number of vilous and crypt cells (Cummins & Thompson, 2002). It is possible that this accelerated epithelial growth is an important host factor in allowing BoNT intoxication by P-47 family encoding strains.

An association between OrfX strains and infant botulism can also be seen in other samples for which the necessary data is available (Fillo et al., 2011). It should be noted that while Fillo et al. did not experimentally determine the toxin complex type, for the purposes of this analysis it was inferred from the toxin type. This inference should be accompanied by the caveat that rare exceptions from the general pattern have been noted (e.g. BoNT/A1 associated with OrfX gene cluster). In their study, there were 9 clinical strains which encoded toxin types typically associated with OrfX gene clusters, 6 of these strains were associated with infant botulism. The other 3 strains were associated with food botulism. One of these strains was bivalent, encoding A2 (OrfX associated) and B1 (HA associated) leaving two clinical strains that are likely to encode only OrfX associated with food-borne botulism, compared with 6 probable OrfX strains associated with infant botulism. In comparison to the 6 infant botulism cases caused by toxin types associated with OrfX encoding strains there were 18 infant botulism cases caused by toxin types associated with HA clusters. Despite the majority of probable OrfX strains investigated by Fillo et al. being associated with infant botulism, it should be noted that BoNT/A2 (OrfX associated) has been linked with outbreaks of food botulism in Italy, so it could be more prevelant in the environment, and hence more likely to cause both food and infant botulism in Italy (Bigliardi & Sansebastianio, 2007). Further characterisation and epidemiological studies are required. Fillo et al. are unique in literature for providing toxin subtype along with tyoe disease caused. The lack of BoNT/A subtyping information and/or information on type of disease

hampers the wider investigation of the hypotheses presented here. For example, of 366 cases of infant botulism diagnosed in Argentina, 100% were BoNT/A, however subtype information is not given (Koepke et al., 2007). Similarly, it is known that at least 39 *C. botulinum* strains encoding BoNT/A subtypes associated with OrfX have been isolated in Argentina but the type of botulism associated with these strains is unknown (Luquez et al., 2012). Between 2001-2010, 52% of infant botulism cases in the United States were due to serotype A producing strains, however, the subtype is not available (http://www.cdc.gov/nationalsurveillance/botulism_surveillance.html). Similarly, while it is known that 50% of food botulism cases in the USA between 1990-2000 were associated with BoNT/A, the proportion caused by different subtypes is not known (Sobel et al., 2004).

While the work of Fillo et al. provides more evidence that OrfX encoding strains are more likely to be associated with infant botulism, it sheds no light on whether this is because the OrfX toxin complex provides an advantage in infant botulism or whether encoding the OrfX toxin complex is a disadvantage to the organism compared to encoding the HA toxin complex in causing food or wound botulism. Due to the lack of functional characterisation of the OrfX proteins we cannot usefully speculate on their role in infant botulism. In the case of food botulism, the possible disadvantage of OrfX compared with HA is that there is evidence that the binding of OrfX to BoNT/A-NTNH is weaker than that of HA to BoNT/NTNH (Lin et al., 2010), however spontaneous association between BoNT/E and OrfX proteins has been reported (Kukreja & Singh, 2007). As larger toxin complexes provide more protection from acids and proteases in the gut (Sakaguchi, 1982) and the HA complex plays a key role in BoNT crossing the gut epithelium (Sugawara et al., 2010; Lee et al., 2012), if OrfX doesn't form a complex with BoNT/NTNH then less

BoNT may enter the circulation. This disadvantage would lessen in the milder conditions of the infant gut, resulting in more OrfX encoding strains causing infant botulism. It is not expected that toxin complex type would have any significant impact on wound botulism toxico-infection as the toxin complex dissociates at physiological pH (Simpson, 2004).

While the presence of OrfX in the toxigenic *C. butyricum* investigated in this work was not experimentally confirmed, BoNT/E is typically associated with OrfX (Hill & Smith, 2013). There were two incidents of botulism associated with the toxigenic *C. butyricum* investigated here, both were cases of infant botulism.

Although the epidemiological evidence suggests that OrfX encoding strains are associated with infant botulism, the lack of a functional explanation that explains this association weakens this hypothesis. Further work on the functional significance of the OrfX proteins in a model of infant botulism would address this hypothesis more fully. Characterisation of the toxin complex type associated with all types of botulism would provide epidemiological evidence that would also address this hypothesis.

4.3.2. Hypothesis 2 – ability to cause different disease is caused by factors present in the accessory genome

An alternative hypothesis to explain the results seen here is that, there is variation between the accessory genomes of the different clusters that is important in determining the type of botulism caused. For example, the genomes of strains in clusters 1 and 2 could contain a virulence factor that enables them to cause wound botulism at a higher rate than other clusters. In other species of bacteria,

virulence factors suggested to enable wound infections include hemolysins in *Aeromonas hydrophila* (Gold & Salit, 1993) and quorum sensing in *Pseudomonas aeruginosa* (Rumbaugh et al., 1999). Likewise, the genomes of strains in cluster 4 could encode a virulence factor that results in them being well suited to cause infant botulism such as pili or other attachment factors, which are known to have been acquired by Gram positive pathogens via horizontal gene transfer (Telford et al., 2006).

In this scenario, differences in disease type are independent of toxin complex type, with the correlation between the two being secondary to the genomic background. This hypothesis of uncharacterised virulence factors resulting in 'specialisation' of different clusters of *C. botulinum* at causing different types of botulism could be investigated using a Genome Wide Association Study (Falush & Bowden, 2006). This approach required the whole genome sequences of a large number of wound, food and infant botulism strains.

4.3.3. Hypothesis 3 – the correlation between disease type and toxin complex type/genomic similarity is due to common exposure

Another hypothesis that could explain the results presented here is that the toxin complex type has no influence on disease type. Alternatively, the correlation between genomic similarity/toxin complex type and disease type could reflect a common geographic source of *C. botulinum* associated with different forms of botulism. It is known that certain toxin types predominate in different regions; a study on *C. botulinum* in British soil identified only BoNT/B producing organisms (Smith & Young, 1980), in the western USA most strains are type A, while east of

the Mississippi River is mostly type B strains (Smith, 1978) and in Argentina the majority of strains are type A (Luquez et al., 2005).

This hypothesis states that, if exposure were the main factor in occurrence of botulism (rather than organism specific factors, as hypothesised above), then clinical cases of *C. botulinum* would follow a similar distribution to those in environment from which they originated, both in terms of toxin type and genomic similarity (i.e. fAFLP clustering). For example, strains associated with wound botulism originate from the country of origin of the heroin. If toxin type and genomic similarity information was available for environmental isolates from these regions, the hypothesis that strains cluster 1 and 2 originate from the same country as the heroin the drug users were taking could be tested. The fact that cluster 1 and cluster 2 are different, and yet both contain wound botulism strains could be explained by different countries of origin for the heroin associated with these two clusters e.g. Afghanistan and Morocco. This theory, of spores in a sample reflecting the origin of the sample, underpins forensic palynology (the use of pollen and spores in solving legal issues), which has been successfully employed for over 50 years (Mildenhall et al., 2006). It is expected that infant botulism strains would be derived from either the local environment or the origin of an implicated foodstuff such as honey or chamomile tea (Luquez et al., 2012) or other source of spores such as pet terrapins. In order for this hypothesis to be accurate, the cluster 4 infant botulism cases would have to be caused by strains from the same region. This region could be the region of exposure, i.e. the UK or the region of origin of a common risk factor i.e. honey. However, the cases in cluster 4 do not share a common risk factor. Food botulism associated strains would, depending on the time of contamination, originate from the place the foodstuff was grown, processed or consumed.

A global phylogeography of *C. botulinum* would allow us at whether phylogeographic source attribution is possible e.g. an infant botulism strain from a major honey producing nation could be linked to honey consumption. Having the whole genome sequence of these strains and a comprehensive global collection of *C. botulinum* would allow these phylogeographic relationships to be explored and directly address this hypothesis.

4.3.4. Hypothesis 4 – differences between toxin types are responsible for disease type specificity

The final hypothesis that should be addressed is that it is differences between the toxin subtypes, rather than the toxin complex, that is responsible for Cluster 4 being associated with infant botulism. This hypothesis assumes that, as is the case for all previously characterised strains, that OrfX strains encoding BoNT/A encode subtypes A2, A3 or A4.

There have been few biological differences identified between the BoNT/A subtypes; BoNT/A2 has been shown to enter neuronal cells faster than BoNT/A1 (Pier et al., 2011) while BoNT/A3 showed intoxication signs significantly different from those seen with BoNT/A1 when injected intravenously into mice (Tepp et al., 2012). Recent work by Whitemarsh et al. (2013) showed that BoNT/A1-A5 have distinct characteristics. Particularly relevant for this work is the fact that when mice were treated intravenously with BoNT/A1-A5, the BoNT/A1 and A5 symptoms closely resembled each other (classic ruffled fur, flaccid paralysis, laboured breathing), while the BoNT/A2 and A3 symptoms resembled each other (paralysis of front legs, hind legs and then whole body). This suggests that, perhaps, BoNT/A1 and A5 have distinct somatic targets or activities compared with

BoNT/A2 and A3. The potential impact of these differences on type of botulism is unclear, but is worthy of consideration in any future work.

4.3.5. Novel toxin and toxin complex combinations

There were two strains identified that had interesting combinations of toxin and toxin complex; H091640054 encoded BoNT/A and B but only the HA complex type and H094460264 which encoded both HA and OrfX toxin complex types, but only BoNT/A.

Three possible toxin and toxin complex arrangements to explain the results seen in H091640054 can be seen in Figure 72. Previous work on characterisation of bivalent A(B) strains has not identified arrangement (A) (Hill et al., 2009; Smith et al., 2007), while arrangement (B) is similar to that observed in the A5(B) strains (Carter et al., 2010), however it could not be an identical arrangement as the *bont/B* gene identified by Carter et al was not detected by the qPCR used here for toxin identification due to a gene truncation, while the same qPCR assay identified the *bont/B* gene here. Arrangement (C) has also not been previously reported in the literature on bivalent A(B) strains (Hill et al., 2009; Smith et al., 2007). Any of these arrangements are feasible explanations for the results obtained in this study. Additionally, the *bont/A* gene could be present in an OrfX complex if there is sequence variation between the PCR primers and the OrfX target sequences that lead to no OrfX product. It would be enlightening to know the toxin sub-type, especially of the *bont/A*.

The second interesting strain, H094460264 was an infant botulism associated strain that encoded genes from two toxin complex types (HA and OrfX) but only

one toxin type, *bont/A*. Possible arrangements that explain this result can be seen in Figure 73. The explanation that fits with our current understanding of the relationship between toxin complex and toxin complex type, is that there are two *bont/A* complex clusters, one of which encodes OrfX while the other encodes HA (Figure 73A). However, there is no evidence presented here that requires both the toxin complexes to be complete and one of them could be missing the *bont/A* gene (Figure 73B & C). Similarly to H091640054, there could also be sequence variation or gene truncation in e.g. *bont/B* that would provide an alternate explanation to the ones listed here.

In both strains detailed above, there is likely to be a novel arrangement of the genes encoding BoNT and the toxin complex. It would be enlightening to obtain the subtype of the *bont* genes encoded by these strains, however, a definitive answer could be obtained from the whole genome sequence of these strains.

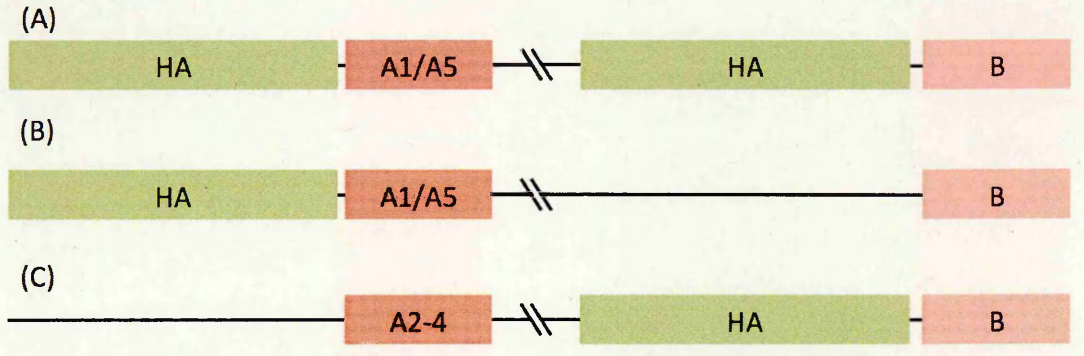


Figure 72: Schematic representation of two possible toxin and toxin complex arrangements that explain the results seen in H091640054. Toxin genes in red/pink, complex in green.

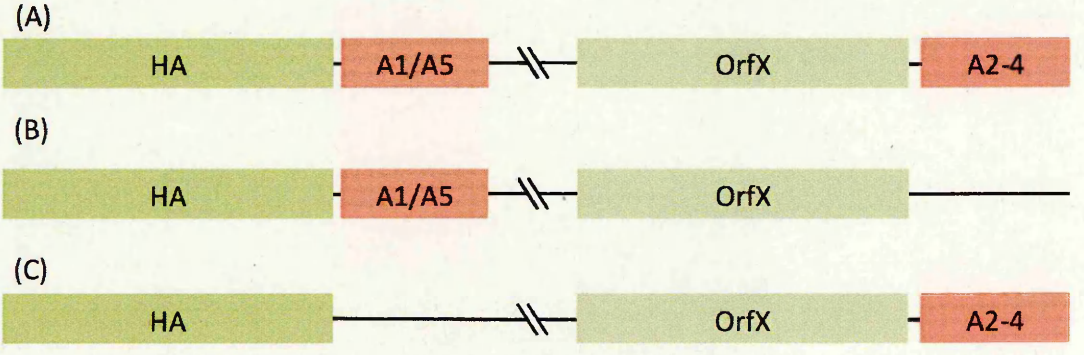


Figure 73: Schematic representation of two possible toxin and toxin complex arrangements that explain the results seen in H094460264

4.4. RNA-Seq transcriptome profiling to investigate C. botulinum toxin complex expression

RNA-sequencing (RNA-seq) is the application of second generation sequencing technologies to the whole transcriptome of an organism. RNA is isolated from an organism, genomic DNA is enzymatically removed and reverse transcription used to create a cDNA library. This cDNA library is then sequenced using the same chemistry as for genomic DNA. Once sequencing results are obtained, the resulting 'reads' are mapped to the genome of the organism from which they were obtained and used to investigate numerous transcriptional phenomena.

4.4.1. Analysis of gene expression data from the RNA-seq dataset

An average of only 11.2% of reads from each sample were mapped to the reference genome. Even considering that with RNA-seq experiments it is only expected to be able to map 40-50% of reads (Blow, 2009) this is low. One possible explanation for this is that a high proportion of reads mapped ambiguously (i.e. to more than one place). In the analysis carried out here, reads that mapped to more than one place were discarded from the analysis. These ambiguous reads include the 16-30% of reads that mapped to 5S, 16S and 23S rRNA. Transcription from other repetitive regions of the genome, such as tRNAs, would also be ambiguously mapped. Another possibility to explain the high proportion of unmapped reads is that the RNA sent for sequencing was contaminated. One way of investigating this would be to use e.g. MEGAN, a metagenome analysis tool, to analyse whether the reads that didn't map to the *C. botulinum* genome were from other organisms. However, this tool requires a level of bioinformatic sophistication that was not available at the time of analysis. Additionally, although ~90% of reads had to be

discarded, the high throughput of the SOLiD system means that an average coverage of the genome of >50x was still achieved with the 11% of reads that mapped. In an analysis by Haas et al. (2012), it was found that reduction in coverage from 25 million to 3.2 million reads had a negligible impact on the proportion of CDSs from which expression was detected. Coverage was more important for the determination of differentially expressed genes, with a 10 fold reduction in number of reads (from 25 million to 2.5 million) resulting in 5-40% fewer differentially expressed genes, depending on the magnitude of the differential expression (genes with higher differential expression were more robust to decreases in coverage) (Haas et al., 2012). This shows that the low proportion of reads mapping to the reference genome in the data presented here may have some negative impact on the number of differentially expressed genes that could be identified.

4.4.2. Quality of RNA-seq data

When validating results obtained from a new methodology it is important to compare with previous gold-standard techniques. In the case of *C. botulinum* transcriptomics, the current gold standard is the microarray. Artin et al., 2010 analysed the transcriptome of *C. botulinum* A ATCC 3502 that is closely related to the ATCC 19397 strains used here. When the normalised expression detected in the Artin et al. microarray experiments was compared with the normalised RNA-seq expression from this work, a correlation similar to previous comparisons of RNA-seq and microarray work was found (Porcelli et al., 2013). The transcriptome of *C. botulinum* is highly active, with 86% of CDSs showing transcription of >10 RPKM.

There were more genes that showed significant differential gene expression (i.e. EdgeR False Discovery Rate of < 0.01) between late-log and early stationary phase than between mid-log and late-log. A higher number of genes being differentially expressed as a bacterial culture enters stationary phase has been observed in *Helicobacter pylori* (Thompson et al., 2003), *Escherichia coli* (Chang et al., 2002) and *Pseudomonas aeruginosa* (Wagner et al., 2003). This is due to widespread physiological responses of the cell to slowing growth rate (Wagner et al., 2003).

4.4.3. RNA-seq analysis of expression from *botA* gene cluster

RNA-seq analysis shows that transcription from the *bont* cluster occurs in two operons; the *bont-ntnH* and *ha33-ha17-ha70*, confirming in ATCC 19397 what had previously been found in strain NCTC 2916 (Henderson et al., 1996; Raffestin et al., 2005).

The concentration of transcripts (i.e. RPKM) from the genes in the two operons (*bont-ntnH* and *ha33-ha17-ha70*) showed a comparable level of similarity to that observed for operons from other Clostridia for which RNA-seq experiments have been carried out (Wang et al., 2012). It is notable however, that between mid-log and late log *botA, ntnH, ha17* and *ha33* increase an average of 4.9 fold while *ha70* increases 6 fold. One hypothesis to explain this is that there is a secondary transcriptional start site within the *ha* operon that results in more *ha70* transcription. Secondary transcriptional start sites have been observed for a number of bacteria, including *Salmonella* Typhimurium and *H. pylori* (Kroger et al., 2012; Porcelli et al., 2013). However, studies that identify secondary transcriptional start sites use an enzyme treatment with terminator exonuclease that degrades non-primary mRNA

(Sharma et al., 2010). This process enriches for the 5' end of mRNA fragments, allowing the single nucleotide identification of transcriptional start sites. Alternatively, for a single transcript, a method such as 5' RACE (rapid amplification of cDNA ends) could be used to obtain the full length sequence of the mRNA (Schramm et al., 2000). It is also clear from the RNA-seq data that clostripain is not co-regulated with the *bont* gene cluster.

A recent paper has determined the crystal structure of the BoNT complex and found the stoichiometry of the different components to be 1:1:3:3:6 (BoNT:NTNH:HA70:HA17:HA33) (Lee et al., 2013). The relatively even transcription from across the BoNT complex cluster indicates that this stoichiometry is determined on a translational level, as has been identified for other toxins such as cholera toxin (Hirst, 1995).

When the expression of the *bont* complex genes is examined it is obvious that *botR* is expressed at much lower levels than *botA, ntnH, ha33, ha17* and *ha70* (i.e. the ANTP genes). BotR is an alternative sigma factor involved in the regulation of the botulinum locus, encoded by the *botR* gene. It is interesting that *botR* is not present in most *botE* encoding strains and is missing it's sigma factor binding site in the *bont/A5* gene cluster without effect (Carter et al., 2010). BotR binds to promoter sequences upstream of *bont* and *ha33* when it is part of a complex with RNA-polymerase (Raffestin et al., 2005). When BotR is overexpressed, there is an increase in production of the BoNT and the ANTPs and their respective mRNAs (Marvaud et al., 1998). Previous work has found that *botR* is expressed at approximately 1% of the level of *botA* (Couesnon et al., 2006). The data in our study finds that *botR* is expressed at 0.6-4% of the level of *botA* and that there is no significant change in the expression of *botR* through the time-course. When this

is compared with the significant upregulation of *botA* and the ANTP encoding genes, the precise role of *botR* in the regulation of the *bont* cluster is unclear. The fact that *C. botulinum* E strains produce toxin and cause disease with no *botR* makes it clear that there are multiple factors involved in the regulation of *bont*. It also seems possible that these other factors play a significant role in the regulation of *botA* in the strain investigated here, due to the lack of any significant change in expression of the *botR* even while *botA* expression increased significantly. In other toxigenic bacteria, small RNAs (sRNAs) are involved in the regulation of virulence factors (Toledo-Arana et al., 2007). These include *C. perfringens*, where an sRNA controls the synthesis of collagenase A and alpha toxin at the transcriptional level (Shimizu et al., 2002).

4.4.4. Problems with using *gluD* as reference gene in RT-qPCR

There is disagreement in the measurement of *botA* expression between the RT-qPCR relative gene expression experiment and the RNA-seq expression analysis. The reason for this discrepancy is apparent if the expression of gluD, the reference gene in the RT-qPCR experiment, is examined. Expression of gluD increased 5.8 fold between late log and early stationary phase. This increase in expression of the reference gene decreased the relative gene expression of *botA*, indicating that the expression of *botA* has decreased (under the assumption that *gluD* expression is stable). This resulted in the sampling of RNA too early i.e. still during high toxin expression rather than post-peak toxin expression. Transcriptome analysis at the time point of post-peak toxin expression was expected to yield interesting information on the expression of the *agr* loci, putative quorum sensing proteins with a potential role in control of sporulation and toxin production (Cooksley et al., 2010) or the expression of sigma factor K which is

purported to be involved in early stage sporulation (Kirk et al., 2012). Unfortunately, due to the upregulation of the reference gene in the relative gene expression experiment, this was not possible.

In order to prevent future relative gene expression experiments from being compromised by poor choice of reference gene, the RNA-seq data was mined for genes that showed stable expression between mid-log, late-log and early stationary growth phases. The genes with the lowest co-efficient of variation (standard deviation divided by mean) across biological replicates are suggested as candidate reference genes for future relative gene expression analyses in *C. botulinum*. These genes include a cell envelope lipoprotein (*clb_0260*), a sporulation protein (*spoVG*), hypothetical proteins and an RNA polymerase sigma factor (*rpoD*). It is expected that transcription of *spoVG* would increase later in the time course as the culture starts to form spores, while it is difficult to predict what changes might occur in the transcription of a hypothetical protein. Therefore if future relative gene experiments were being planned, mRNA from *clb_0260* (lipoprotein) or *rpoD* would be considered as reference genes. Additionally, in this work the TaqMan quantitative PCR system was used which provides excellent specificity, reproducibility and quantification. However, the custom fluorescent oligonucleotide probes necessary for TaqMan assays are expensive, prohibiting the testing of multiple reference genes for stability of expression. In future work, SYBR Green dsDNA-binding dye would be considered, as it is non-specific, allowing the testing of multiple candidate reference genes. The most commonly used reference in relative gene experiments is the *rrn* 16S rRNA locus. This was not used here as changes in the rate of synthesis of rRNA as growth slows have been observed (Hansen et al., 2001) and proposed to interfere with the calculation

of the relative gene expression of *bont* (Lovenklev et al., 2004). Housekeeping genes selected for MLST schemes are often proposed as reference genes for relative gene expression experiments, none of the *C. botulinum* MLST scheme genes were among the 10 genes with the lowest coefficient of variation.

4.4.5. Identification of up-regulated potential pathogenicity genes

The RNA-seq data was mined to identify potential virulence factors with a similar expression profile to *C. botulinum*. There were 6 thermolysin metallopeptidase genes (*npr-1* to *6*) that showed significant upregulation between late log and early stationary phase. The protein product of *npr-6* was identified in the whole supernatant proteome analysis at both 24 h and 96 h and was found to be similar to the *C. perfringens* lambda toxin that activates *C. perfringens* epsilon toxin. Expression of these genes may be under-estimated because the RNA-seq reads were only allowed to map unambiguously and these sequences are around 79% similar to each other. Two other genes that were significantly upregulated in early stationary phase with similarity to previously characterised virulence factors were *clb_1609* and *clb_0527* that encode a hemolysin and a bacteriocin respectively. In *Bacillus cereus*, an enteric pathogen, hemolysin is an exotoxin that effects endothelial integrity (Beecher et al., 1995). Pore forming toxins, such as hemolysins, are also implicated in the disruption of epithelial gap junctions and the necrosis of intestinal epithelium (Los et al., 2013). Hemolysins are typically upregulated in the absence of iron as the organism attempts to scavenge iron from the heme of lysed hemocytes (Griffiths & McClain, 1988). This indicates that as the *C. botulinum* culture is entering stationary phase, nutrients such as iron are becoming limited. It is possible that *C. botulinum* hemolysin plays a role in

disrupting host gut epithelium in food botulism or ensuring an adequate supply of iron for the organism in a wound botulism infection.

4.4.6. Analysis of small RNAs

Second-generation sequencing technologies have led to an explosion in the identification and understanding of regulatory roles for small, non-coding RNAs (sRNAs). In bacteria, sRNAs have been found to control stress response and to regulate expression of virulence factors (Toledo-Arana et al., 2007; Sharma & Heidrich, 2012). Here, we find the transcriptome of *C. botulinum* to be highly active in terms of both characterised sRNAs and uncharacterised sRNAs.

As the focus of this project is the production of the toxin, the transcriptome of *C. botulinum* was examined for potential small RNAs that influence *botA* expression. One common way for sRNAs to influence the expression of a gene is on a post-transcriptional level. The homology-mediated binding of an sRNA to the 5' Untranslated Region (5' UTR) of an mRNA can occlude the ribosome binding site of the mRNA, dramatically slowing it's rate of translation (Papenfort & Vogel, 2010). One piece of evidence that sRNAs could be involved in regulation of *bont* is the intriguing finding that in one bivalent strain that produces BoNT/B and BoNT/F, the relative amounts of toxin produced were temperature dependent (Barash & Arnon, 2004). RNA mechanisms are known to be responsible for temperature dependent regulation of pathogenicity factors, with the canonical example being *prfA* in *Listeria monocytogenes* (Johansson et al., 2002). Additionally growth temperature also effects the patterns of relative expression from the toxin gene cluster in *C. botulinum* producing BoNT/E, with growth at 30°C resulting in more *botE*

transcription compared with growth at 10°C (Chen et al., 2008). All sRNAs identified were examined for homology with the 5' end of the *botA* mRNA and the *ha33* mRNA. No candidates were identified but an active transcriptional landscape of non-coding sRNAs was uncovered.

The database Rfam (Burge et al., 2012) is a collection of non-coding RNA families, many of which have putative functions assigned. When the sRNAs in *C. botulinum* were compared against Rfam using Infernal (Nawrocki et al., 2009) there were 17 transcriptionally active T-boxes. These are found in the 5' UTR of of mRNAs encoding amino-acid tRNA ligases, they upregulate tRNA ligase expression in the presence of a cogent uncharged tRNA, resulting in more enzymes to link amino acids with the appropriate tRNA (Gutierrez-Preciado et al., 2009). Additionally, there were 19 transcriptionally active riboswitches which are 5' UTR elements that sense the presence of small metabolites such as magnesium and purines and altering the regulation of transporters of those metabolites (Bastet et al., 2011). While these sRNAs have been previously characterised, this is first time that they have been experimentally identified in *C. botulinum* and they provide an insight into the amount of sRNA mediated regulation that is taking place in this organism.

Perhaps the most interesting sRNAs are the uncharacterised ones. There was an sRNA, approximately 500 bp long, that showed strong growth phase regulation. The RPKM at mid-log was 551, at late log it was 851 and at late stationary it was 36397. This dramatic increase in expression was not mirrored in either of the flanking genes, which are both on the positive strand while the sRNA is on the negative strand. The genomic region encoding the sRNA is highly conserved (100%) in other sequenced proteolytic *C. botulinum* but shows no significant homology in any other species. The sRNA is part of an approximately 23 kbp long,

highly conserved region of the genome. The function of this sRNA is a mystery, however, it's level of expression is more analogous with structural sRNAs such as 4.5S RNA which is involved in trafficking proteins with signal peptides (Waters & Storz, 2009) rather than the expression levels of regulatory sRNAs. There were also high levels of expression of a large sRNA, the Ornate Large Extremophilic RNA that has been previously identified in extremophilic bacteria (Puerta-Fernandez et al., 2006). It has been posited that this sRNA has a role in protecting *Bacillus halodurans* from alcohol toxicity (Wallace et al., 2012). It's role in *C. botulinum,* where it is expressed at high levels at all time points but with an in intensity that increases through the growth curve is unclear, although perhaps it is a response of the cell to the production of volatile secondary metabolites and by products of fermentation. Finally, there were at least 9 occurrences through the genome of short areas of sense-antisense transcription. Some of these share some characteristics of toxin-antitoxin systems – i.e. one strand encodes short open reading frame with Shine-Delgarno sites, typical of type I toxin-antitoxin systems. However, the primary and secondary structure of the peptides is not typical of type I toxin-antitoxin system which typically have high hydrophobicity and alpha-helices (Fozo et al., 2008). Characterisation of these active transcriptional regions requires further work.

4.5. Final discussion and future work

The botulinum toxin complex in its entirity is important for the toxicity of the botulinum neurotoxin. All *bont* genes characterised to date are co-localised with one of two toxin complex types. The fact that this co-localisation is ever-present, despite evidence of there being extensive recombination in this locus (Hill et al., 2009) suggests that strains where the *bont* is separate from the complex are not

successful. There are two toxin complex arrangements that are associated with disease, the HA and the OrfX, with the majority of disease associated with HA encoding strains (Hill & Smith, 2013).

This work provides further evidence for the importance of the toxin complex proteins through a combination of *in silico*, proteomic and genomic experiments that were directed to achieve the aims laid out in section 1.6. These aims were to use *in silico* techniques to analyse the toxin complex protein sequences, a proteomic investigation of the *C. botulinum* whole supernatant proteome, to chracterise the toxin complex components produced by clinical strains associated with botulism, to explore the association between toxin complex type, genomic similarity and disease type and to analyse the *C. botulinum* transcriptome for insight into the regulation of *bont*.

Prior to this work, the only evidence that P-47 family proteins played a role in toxicity was their proximity to the BoNT encoding gene: there is no evidence that they form a complex with BoNT/A-NTNH (Lin et al., 2010) although a spontaneous association between BoNT/E and the P-47 family proteins has been reported (Kukreja & Singh, 2007). The findings presented in this work on clusters of P-47 family sequences encoded alongside putative toxin sequences provides further evidence that these proteins are likely to contribute to toxicity in various species by having a functional role. Additionally, there are other similarities between some of the putative toxins in P-47 family clusters and BoNT. An example of this is the nematicidal proteins. In addition to being encoded alongside P-47 clusters, these toxins share two other characteristics with BoNT; they are often active via the oral route and they are also expressed as part of large, non-P-47 family protein complexes known as Xpt toxin complexes (Sheets et al., 2011). Thus, nematicidal

proteins like BoNT are associated with two types of toxin complex; P-47 family clusters and distinct non-P-47 family complexes (Xpt and HA complex respectively). It is intriguing that these two oral toxins, BoNT and the nematicidal proteins, are both found as part of P-47 gene clusters and alternative toxin complexes. Our understanding of complex aggregations of toxin associated proteins and their contribution to in vivo pathogenicity is constantly developing. This is very well illustrated by two things; the recent discovery that typhoid toxin has a novel A2B5 stoichiometry (Song et al., 2013) and the modularity of the BoNT-NTNH and HA toxin sub-complexes (Lee et al., 2013). Firstly, the novel stoichiometry of the typhoid toxin consists of a typical AB5 toxin where the activity subunit is closely bound to the binding pentamer with an additional activity subunit joined by a single disulphide bond to the other activity subunit (Song et al., 2013). It has been proposed that this structure constitutes a 'missing link' in toxin evolution, and that in millions of years a better-integrated multicomponent toxin will have replaced the A2B5 toxin (Stebbins, 2013). The finding of Lee et al. (2013), that the BoNT-NTNH and HA proteins form structurally and functionally separate sub-complexes also has implications for our model of the role of the ANTPs. In this light it is tempting to hypothesise that the P-47 family cluster is an evolutionary pre-cursor to the HA sub-complex in the case of BoNT and the Xpt complex for nematicidal proteins.

The fact that, in the limited literature on the subject (Lin et al., 2010), there is no direct interaction between BoNT/A-NTNH and the P-47 family proteins indicates that if the P-47 family proteins play a role in BoNT intoxication, it is via a different model to the HA proteins. However, the fact that there has been spontaneous association reported between BoNT/E and P-47 family proteins (Kukreja & Singh, 2007) highlights the lack of understanding of the interaction between these

proteins. There may be an interesting parallel with the bi-modular structure identified by Lee et al. (2013). One potential model is that in the P-47 family encoding strains, the BoNT-NTNH forms the tightly bound complex that has been previously identified (Gu et al., 2012) while the P-47 family proteins form a separate complex that fulfils a similar role to the HA sub-complex i.e. the disruption of the gut epithelium (Sugawara et al., 2010; Lee et al., 2013). The interface between the HA sub-complex and the BoNT-NTNH sub-complex is small, and their structures and functions are quite separate. Perhaps in the P-47 family encoding strains there is no link between the BoNT-NTNH sub-complex and the P-47 family complex, but the P-47 family protein complex is still responsible for disruption of the gut epithelium. This hypothetical arrangement of separate sub-complexes could be less efficient at translocating BoNT across the gut epithelium than the HA sub-complex. This could be because, either the P-47 family complex is not as efficacious at epithelium disruption as the HA complex or the gut epithelium disrupted by the P-47 family complex and the BoNT-NTNH are not co-localised. In this scenario, the linked BoNT-NTNH-HA sub-complex arrangement would have an evolutionary advantage, being more efficient at intoxication, resulting in the BoNT-NTNH-HA complex arising and causing more disease.

Previous work looking at the expression of the toxin complex proteins has always focused on a small number of type strains (Bradshaw et al., 2004; Cheng et al., 2008; Tepp et al., 2012; Jacobson et al., 2011; Lin et al., 2010). This is the first study to look at the expression of the toxin in a larger number of clinical strains. This kind of broad investigation is required to get a fuller picture of the importance of the toxin complex to the virulence of the organism. The expression of at least one toxin complex protein in every strain and the production of all toxin complex proteins by the majority of strains indicates the importance of the toxin complex in

the pathogenicity of the organism. There were only three P-47 family protein-producing strains analysed here. None of them produced all the components of the toxin complex and OrfX1 and P-47 were not identified in any of the strains. It should be noted that two of the OrfX producing strains only had targeted bands of the SDS-PAGE gel analysed by LC-MS/MS and that an analysis of the full supernatant proteome would be informative. Previous work has also identified only some of the P-47 family proteins in *C. botulinum* A2 (Lin et al., 2010). Without the functional characterisation of P-47 family proteins it is not possible to speculate on implications of partial expression. For example, if there were a large amount of functional overlap between the different proteins, expression from one loci could confer the same phenotypic effect as expression of all loci. Probing the details and the regulation of expression of BoNT/A and the HA proteins from NCTC 2916, when the HA proteins are in one toxin gene cluster and BoNT is expressed from another gene cluster would provide insight into the control of expression of the toxin and toxin complex genes. RT-qPCR experiments could be used to assess whether this regulation occurs on a transcriptional or translational level i.e. are *bont/B* and the *orfX* genes transcribed into mRNA or not? Is expression of the *bont/A* and *ha* genes coordinated to the same degree as strains where these genes are encoded in the same toxin gene cluster? This would be a significant improvement on previous work on this subject that has been done using northern blots (Bradshaw et al., 2004).

The expression of the toxin complex proteins, which are among the most metabolically expensive proteins in the *C. botulinum* supernatant, indicates that they are important to the ecological niche of the organism. There appears to be a relationship between extracellular protein cost and virulence in other organisms, including *S. aureus* and *L. monocytogenes* (Burlack et al., 2007; Pocsfalvi et al.,

2008; Dumas et al., 2009). This novel way of identifying potential virulence factors builds on previous work by Smith & Chapman, (2010) and the solid evolutionary framework they describe. The poor correlation between predicted supernatant proteins and empirically identified supernatant proteins suggests that these tools are of limited utility.

The large number of supernatant proteins identified in this work indicates that the *C. botulinum* supernatant proteome is more complex than previously thought (Cheng et al., 2008). The impact of this complexity on virulence is uncertain, although the identification of potential virulence factors in the supernatant proteome suggests that this warrants further investigation. Ideally, the intra-gastric lethality tests of Cheng et al., (2008) would be repeated with the culture supernatant from this work in order to determine whether any of the potential novel virulence factors identified here have an impact on pathogenesis compared with the purified toxin. The post-translational nicking of HA70 identified in this study has been proposed to allow the formation of the HA70 trimer that mediates the binding of the HA sub-complex, with the BoNT-NTNH sub-complex (Lee et al., 2013). This indicates that the BoNT complex identified in this study is in the form suggested by Lee et al. (2013). Also, the results presented here suggest that the cost of extracellular proteins associated with virulence is higher than is typical for extracellular proteins. This finding was replicated in a range of other bacterial pathogens for which the extracellular proteome had been characterized (Burlack et al., 2007; Pocsfalvi et al., 2008; Dumas et al., 2009; Kaakoush et al., 2010).

There were four hypotheses generated to explain the results from the comparison of fAFLP, toxin complex type and type of disease caused. Reflecting on the results

presented in the fAFLP section, the hypothesis that P-47 family proteins provide a functional advantage in the causation of infant botulism requires significant further work. Understanding of the nature of this putative advantage could also be increased by investigating the histological differences between the infant and adult gut epithelium, perhaps with a particular focus on the epithelium of the weaning infant. Previously, mouse models have been useful in the identification of age-dependent host factors that influence susceptibility to infectious disease (Pott et al., 2012). If the association between toxin type and disease type is valid then this is very strong evidence of the importance of the toxin complex. The hypothesis that suggests that uncharacterised virulence factors in fAFLP cluster 4 are responsible for the association with infant botulism is the weakest presented here. Hoping to find 'virulence factors' that explain differences between strains is rather speculative although it would still be useful to address this hypothesis. One piece of evidence in favour of this hypothesis is the close genomic relationship of BoNT/A2 and A3 strains as assessed by MLST and PFGE (Luquez et al., 2012), indicating that they share a common evolutionary history and could have a common set of pathogenicity factors. The fourth hypothesis; that functional differences in activity between different subtypes of BoNT/A could lead to differences in type of disease caused, arises largely due to the intriguing reslts of Whitemarsh et al., (2013). When they compared the *in vivo* symptoms of BoNT/A1-A5, they found that BoNT/A1 and A5 had shared characteristics while BoNT/A2 and A3 had distinct shared characteristics. This is relevant as the former are associated with HA toxin complex and the later with P-47 family cluster and so cluster 4 could contain BoNT/A2 and A3 strains. Therefore, cluster 4 could be associated with infant botulism due to a shared functional characteristic such as the one identified by Whitemarsh et al. (2013). However, the relevance of their findings to food, infant and wound botulism is uncertain as they intoxicated their

mice with large doses of toxin delivered intravenously. It would help to address this hypothesis if the toxin subtype were determined for the BoNT/A strains as if the cluster 4 strains encoded BoNT/A4 then support for this hypothesis would diminish.

The current understanding of infant botulism fits with the hypothesis that exposure to spores is a key determinant in the development of *C. botulinum*. This understanding is that infant botulism is a disease caused by a combination of a susceptible host (due to a disrupted gastrointestinal flora) and environmental exposure, with not much consideration given to the role of pathogen factors (Koepke et al., 2007; Fenicia & Anniballi, 2009). However, given the small amount of global context for the isolates investigated here, it is premature to dismiss the alternative hypotheses. Of course, the relationship between OrfX and infant botulism identified here could be insignificant, further epidemiological studies are required to confirm or refute this relationship, although studies not designed with this aim in mind provide additional support (Fillo et al., 2011). As infant botulism is thought to be associated with a disrupted gut microbiota, an interesting experiment would be to compare the gut flora of infants diagnosed with infant botulism compared with those that have not developed botulism. This could be done using faecal samples sent to the Botulinum Reference Laboratory, as long as the samples were taken prior to antibiotic treatment. Either shotgun metagenomics or bacterial community profiling (i.e. using 16S rDNA) could be used to profile the level of diversity in the gut flora of infants with and without botulism to determine if there is an association between complexity of gut flora and susceptibility to infant botulism. Additionally, an *in vitro* model of the infant gut could be used and seeded with different levels of bacterial diversity (Nicholson et al., 2012). The ability of *C. botulinum* spores to colonise this environment could then be assessed.

It is difficult to compare the strains investigated here to previously investigated strains due to differences in the enzymes used for fragmentation of the DNA in the fAFLP analysis. The use of a sequence based method, such as whole genome sequencing, would create a portable resource that could also inform future studies of diversity. The evidence presented here, that of the 5 OrfX encoding strains, all 5 were associated with infant botulism warrants further investigation. It would be very informative to know the number of cases of infant botulism caused by OrfX encoding strains, compared with the number of cases of all forms of botulism caused by OrfX encoding strains. However, the information currently in the literature only allows a piecemeal analysis. If the relationship between OrfX and infant botulism is borne out, it is likely to be the result of a combination of factors. The evidence presented here suggests that one factor worthy of investigation is the role of the P-47 family proteins.

Another question that should be addressed is whether the Botulinum Reference Laboratory should routinely determine toxin complex type. While it would not effect clinical management of the cases in the way that knowledge of the toxin type does (appropriate anti-toxin can be given). It would help confirm or deny epidemiological questions about the association between infant botulism and OrfX encoding strains. This epidemiological support would help justify functional studies that could lead to an increased understanding of the route of infection.

There were two strains with interesting, potentially novel combinations of toxin and toxin complex genes. Obtaining the whole genome sequence of these isolates would provide a definitive answer as to the toxin gene arrangement in these strains. Increasing our understanding of the diversity of toxin gene arrangements,

especially in bivalent strains, may provide insight into the co-evolution of the toxin and toxin complex.

Due to flaws in the experimental design (i.e. inappropriate choice of reference gene) the RNA-seq work presented here has not been particularly useful to directly address the importance of the toxin complex to the development of botulism. However, the RNA-seq work has found that the expression of *botR* undergoes no significant detectable change between mid-log and early stationary phase, despite the significant increase in *bont* expression. Furthermore, the RNA-seq data has raised questions as to the role of small RNAs in the regulatory and functional landscape of *C. botulinum* that it would be very interesting to probe further as the number of sRNAs involved in expression of bacterial pathogenicity factors continues to increase (Loh et al., 2013).

This work fits well into the existing *C. botulinum* and BoNT literature. It builds on the increasing amount of literature that supports the view that the botulinum toxin complex is important for disease (Lee et al., 2013; Sugawara et al., 2010) but from an 'omic' point of view, compared with previous functional studies. These two approaches to science are complementary, and a significant portion of the future work to test the hypotheses presented here is more functional in nature i.e. the impact of knock-outs on BoNT production and toxicity. There has not been a lot of omic work performed on *C. botulinum* and where there has, this study has aimed to build on that work and use the latest technologies to provide new insight. For example, Artin et al. (2010) performed enlightening transcriptomic work on proteolytic *C. botulinum* using microarrays and the work performed here with RNA-seq provides a different perscpective to that work, allowing characterisation of novel features such as sRNAs.

A full characterisation of the transcriptional activity and expression of proteins from the P-47 family cluster in *C. botulinum* would be a suitable first step. Investigating the interaction between the P-47 family proteins and gut epithelial cells would be an interesting additional experiment to probe their possible function. Establishing the crystal structure of these proteins would enable more specific hypotheses to be generated as to their function. As part of these experiments it would also be useful to establish the crystal structure of the NTNH associated with the P-47 family cluster. This is because the OrfX associated NTNH lacks a motif that mediates the interaction between NTNH and the HA sub-complex (Lee et al., 2013). The crystal structure of the OrfX NTNH would allow the investigation of the hypothesis that there is an alternative motif that mediates interaction with P-47 family proteins. Additionally, knocking-out the P-47 family encoding genes using e.g. ClosTron (Heap et al., 2007) and examining the effect on toxicity in an animal model would help address the question of P-47 family protein function.

Obtaining the whole genome sequence of a representative selection of *C. botulinum* isolates associated with different diseases and different geographical regions would help to address the hypotheses proposed by the fAFLP work. It would also help address questions about the evolution of the toxin, the toxin complex and the organism. For example, is OrfX the ancestral toxin complex? Finally, functional studies looking at the phenotype of strains that had the growth phase dependent sRNA knocked out would elucidate whether there was any involvement between this and toxin production or sporulation.

Originally the role of the HA proteins was thought to be relatively minor, only providing protection against degradation in the gut. However, recent work has

shown that the HA forms a sub-complex that is vital for the toxin to translocate the gut epithelium (Sugawara et al., 2010; Lee et al., 2013). Although the role of the P-47 family proteins is currently obscure, the results presented here provide encouragement that further work may result in the uncovering of interesting, novel and widely used mechanisms by which a toxin complex can enable toxin activity.

5. Abbreviations

AB toxin = Activity Binding toxin

AFLP = Amplified Fragment Length Polymorphism

ANTPs = Associated Non-Toxic Proteins

BIG-IV = Botulism Immune Globulin Intravenous

BoNT = Botulinum Neurotoxin

BoNT Hc/Lc = BoNT heavy chain/light chain

bp = base pair

CBI = C. botulinum isolating medium

CDS = coding sequence

DTT = dithiothretiol

ELISA = Enzyme Linked Immunosorbent Assay

EF = Anthrax toxin edema factor

fAFLP = fluorescent Amplified Fragment Length Polymorphism

FB = food botulism

gDNA = genomic DNA

IB = infant botulisum

IDUs = injecting Drug Users

kDa = kilodalton

LF = Anthrax toxin lethal factor

MBA = Mouse Bioassay

mRNA = messenger RNA

ORF = Open Reading Frame

PA = Anthrax toxin protective antigen

PFGE = Pulse Field Gel Electrophoresis

PTX = Pertussis toxin

qPCR = Quantitative PCR

rDNA = ribosomal DNA

RME = Receptor mediated endocytosis

rRNA = ribosomal RNA

RT = room temperature

RT-qPCR = reverse transcripton quantitative PCR

SDW = sterile distilled water

SIDS = Sudden Infant Death Syndrome

SNARE = Soluble NSF (N-ethylmaleimide sensitive fustion protein) Attachment Protein Receptor

SV = Synaptic Vesicle

SV2 = Synaptic Vesicle protein 2 aka synaptotagmin

TeNT = Tetanus Neurotoxin

TER = trans-epithelial electrical resistance

VAMP = vesicle associated membrane protein aka synaptobrevin

WB = wound botulism

Xpt = Xenorhabdus protein toxin

6. Reference List

Allen, S, C Emery, and D Lyerly. "*Clostridium*." 8th ed. Manual of Clinical Microbiology 2003. 835-56.

Akbulut, D., Grant, K. a & McLauchlin, J., 2004. Development and application of Real-Time PCR assays to detect fragments of the Clostridium botulinum types A, B, and E neurotoxin genes for investigation of human foodborne and infant botulism. *Foodborne pathogens and disease*, 1(4), pp.247–57.

Akbulut, D. et al., 2005. Wound botulism in injectors of drugs: upsurge in cases in England during 2004. *Eurosurveillance*, 10(9).

Arndt, J.W. et al., 2005. The structure of the neurotoxin-associated protein HA33/A from Clostridium botulinum suggests a reoccurring beta-trefoil fold in the progenitor toxin complex. *Journal of molecular biology*, 346(4), pp.1083–93.

Arnon, S.S. et al., 2001. Botulinum Toxin as a Biological Weapon. *JAMA*, 285(8), pp.1059–1070.

Arnon, S.S. et al., 1978. Intestinal Infection and Toxin Production by Clostridium botulinum as one cause of Sudden Infant Death Syndrome. *The Lancet*, 311(8077), pp.1273–1277.

Arnon, S.S., 2007. Creation and Development of the Public Service Orphan Drug Human Botulism Immune Globulin. *Pediatrics*.

Artin, I. et al., 2008. Effects of Carbon Dioxide on Neurotoxin Gene Expression in Nonproteolytic Clostridium botulinum Type E. *Applied and Environmental Microbiology*, 74(8), pp.2391–2397.

Artin, I. et al., 2010. Effects of Carbon Dioxide on Growth of Proteolytic Clostridium botulinum , Its Ability To Produce Neurotoxin, and Its Transcriptome. *Applied and Environmental Microbiology*, 76(4), pp.1168–1172.

Ba-Thein, W. et al., 1996. The virR / virS Locus Regulates the Transcription of Genes Encoding Extracellular Toxin Production in Clostridium perfringens. *Journal of Bacteriology*, 178(9), pp.2514–2520.

Banu, S. et al., 2000. Identification of novel VirR/VirS-regulated genes in Clostridium perfringens. *Molecular microbiology*, 35(4), pp.854–64.

Bastet, L. et al., 2011. New insights into riboswitch regulation mechanisms. *Molecular microbiology*, 80(5), pp.1148–54.

Bendtsen, J.D. et al., 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology*, 340(4), pp.783–95.

Bendtsen, J.D. et al., 2005. Non-classical protein secretion in bacteria. *BMC microbiology*, 5, p.58.

Barash, J. & Arnon, S., 2004. Dual Toxin-Producing Strain of Clostridium botulinum Type Bf Isolated from a California Patient with Infant Botulism. *Journal of clinical microbiology*, 42(4), pp.1–4.

Barash, J.R. & Arnon, S.S., 2013. A Novel Strain of Clostridium botulinum That Produces Type B and Type H Botulinum Toxins. *The Journal of infectious diseases*, pp.1–9

Barth, H. et al., 2004. Binary Bacterial Toxins: Biochemistry , Biology , and Applications of Common Clostridium and Bacillus Proteins. , 68(3), pp.373–402.

Beecher, D., Schoeni, J. & Wong, A., 1995. Enterotoxic activity of hemolysin BL from Bacillus cereus. *Infection and Immunity*, 63(11), p.4423.

Bigliardi, L. & Sansebastiano, G., 2007. An outbreak of botulism in Italy. In *Case Studies in Food Safety and Environmental Health*. pp. 57–60.

Blasi, J. et al., 1993. Botulinum neurotoxin A selectively cleaves the synaptic protein SNAP-25. *Nature*, 365, pp.160–63.

Blow, N., 2009. The digital generation. *Nature*, 458, pp.239–242.

Bonventre, P.F. & Kempe, L.L., 1960. Physiology of toxin production by Clostridium botulinum types A and B. Growth, autolysis, and toxin production. *Journal of bacteriology*, 79, pp.18–23.

Bowen, D. et al., 1998. Insecticidal Toxins from the Bacterium Photorhabdus luminescens. *Science*, 280(5372), pp.2129–2132.

Bradshaw, M. et al., 2004. Regulation of neurotoxin complex expression in Clostridium. *Anaerobe*, 10, pp.321–333.

Brett, M.M., Hallas, G. & Mpamugo, O., 2004. Wound botulism in the UK and Ireland. *Journal of Medical Microbiology*, 53(6), pp.555–561.

Browning, L.M. et al., 2011. An outbreak of food-borne botulism in Scotland, United Kingdom, November 2011. *Euro surveillance : bulletin Européen sur les maladies transmissibles = European communicable disease bulletin*, 16(49), p.20036.

Brunger, A.T. & Rummel, A., 2009. Receptor and substrate interactions of clostridial neurotoxins. *Toxicon*, 54(5), pp.550–60.

Bumann, D., Aksu, S. & Wendland, M., 2002. Proteome Analysis of Secreted Proteins of the Gastric Pathogen Helicobacter pylori. *Infection and Immunity*, 70(7), pp.3396–3403.

Burge, S.W. et al., 2013. Rfam 11.0: 10 years of RNA families. *Nucleic acids research*, 41(Database issue), pp.D226–32.

Burlak, C. et al., 2007. Global analysis of community-associated methicillin-resistant Staphylococcus aureus exoproteins reveals molecules produced in vitro and during infection. *Cellular microbiology*, 9(5), pp.1172–90.

Byard, R. et al., 1992. Clostiridum botulinum and sudden infant death syndrome: a 10 year prospective study. *Journal of Paediatrics and Child Health*, 28(2), pp.156–7.

Carter, A.T. et al., 2010. Further characterisation of proteolytic Clostridium botulinum type A5 reveals that neurotoxin formation is unaffected by loss of the cntR (botR) promoter sigma factor binding site. *J. Clin. Microbiol.*, 48(3), pp.1012–3.

Carter, A.T. et al., 2013. The type F6 neurotoxin gene cluster locus of group II clostridium botulinum has evolved by successive disruption of two different ancestral precursors. *Genome biology and evolution*, 5(5), pp.1032–7.

Chang, D.-E., Smalley, D.J. & Conway, T., 2002. Gene expression profiling of Escherichia coli growth transitions: an expanded stringent response model. *Molecular microbiology*, 45(2), pp.289–306.

Chatzi, K.E. et al., 2013. Breaking on through to the other side: protein export through the bacterial Sec system. *The Biochemical journal*, 449(1), pp.25–37.

Chen, Y. et al., 2008. Quantitative Real-Time Reverse Transcription-PCR Analysis Reveals Stable and Prolonged Neurotoxin Cluster Gene Activity in a Clostridium botulinum Type E Strain at Refrigeration Temperature . *Appl. Environ. Microbiol*, 74(19), pp.6132–6137.

Cheng, L.W. et al., 2008. Effects of purification on the bioavailability of botulinum neurotoxin type A. *Toxicology*, 249, pp.123–129.

Cobos, R. et al., 2010. Cytoplasmic- and extracellular-proteome analysis of Diplodia seriata: a phytopathogenic fungus involved in grapevine decline. *Proteome science*, 8, p.46.

Collins, M.D. et al., 1994. The Phylogeny of the Genus Clostridium : Proposal of Five New Genera and Eleven New Species Combinations. *International Journal of Systematic Bacteriology*, (0734).

Collins, M.D. & East, A.K., 1998. Phylogeny and taxonomy of the food-borne pathogen Clostridium botulinum and its neurotoxins. *Journal of Applied Microbiology*, 84(1), pp.5–17.

Connan, C. et al., 2012. Two-Component Systems Are Involved in the Regulation of Botulinum Neurotoxin Synthesis in Clostridium botulinum Type A Strain Hall. *PloS one*, 7(7), p.e41848.

Cooksley, C.M. et al., 2010. Regulation of Neurotoxin Production and Sporulation by a Putative agrBD Signaling System in Proteolytic Clostridium botulinum. *Applied and Environmental Microbiology*, 76(13), pp.4448–4460.

Couesnon, A., Raffestin, S. & Popoff, M.R., 2006. Expression of botulinum neurotoxins A and E , and associated non-toxin genes , during the transition phase and stability at high temperature : analysis by quantitative reverse transcription-PCR. *Microbiology*, 152, pp.759–770.

Couesnon, A., Pereira, Y. & Popoff, M.R., 2008. Receptor-mediated transcytosis of botulinum neurotoxin A through intestinal cell monolayers. *Cellular Microbiology*, 10(September 2007), pp.375–387.

Couesnon, A., Shimizu, T. & Popoff, M.R., 2009. Differential entry of botulinum neurotoxin A into neuronal and intestinal cells. *Cellular microbiology*, 11(2), pp.289–308.

Critchley, E.M., 1991. A comparison of human and animal botulism: a review. *Journal of the Royal Society of Medicine*, 84(5), pp.295–8.

Cummins, A.G. & Thompson, F.M., 2002. Effect of breast milk and weaning on epithelial growth of the small intestine in humans. *Gut*, 51(5), pp.748–754.

DasGupta, B.R., 2006. Botulinum neurotoxins: perspective on their existence and as polyproteins harboring viral proteases. *The Journal of general and applied microbiology*, 52(1), pp.1–8.

Davis, J., Mattman, L. & Wiley, M., 1951. Clostridium botulinum in a fatal wound infection. *The Journal of the American Medical Association*, 146(7), pp.646–648.

Degeneres, L.A., 2008. Waterfowl toxicology: a review. *Veterinary Clinics of North America Exotic Animal Practices*, 11, pp.283–300.

Dekleva, M.L. & Dasgupta, B.R., 1990. Purification and characterization of a protease from Clostridium botulinum type A that nicks single-chain type A botulinum neurotoxin into the di-chain form. *Journal of bacteriology*, 172(5), pp.2498–503.

Dineen, S.S. et al., 2004. Nucleotide sequence and transcriptional analysis of the type A2 neurotoxin gene cluster in Clostridium botulinum. *FEMS microbiology letters*, 235(1), pp.9–16.

Dodds, K., 1992a. Epidemiology of human foodbourne botulism. In A. Hauschild *Clostridium botulinum: Ecology and control in food*. New York: Marcel Dekker Inc, pp. 69-104.

Dodds, K., 1992b. Worldwide Incidence and Ecology of Infant Botulsim. In A. Hauschild, ed. *Clostridium botulinum: Ecology and control in food*. New York,: Marcel Dekker Inc, pp. 105-117.

Dodds, K., 1992c. Clostridium botulinum in the environment. In A. Hauschild, ed. *Clostridium botulinum: Ecology and control in food*. New York: Marcel Dekker Inc, pp. 21-53.

Dong, M. et al., 2003. Synaptotagmins I and II mediate entry of botulinum neurotoxin B into cells. *The Journal of cell biology*, 162(7), pp.1293–303.

Dong, M. et al., 2006. SV2 is the protein receptor for bont. *Science*, 312, pp.592–96.

Dover, N. et al., 2013. Molecular Characterization of a Novel Botulinum Neurotoxin Type H Gene. *The Journal of infectious diseases*, pp.1–11.

Doxey, A.C. et al., 2008. Insights into the evolutionary origins of clostridial neurotoxins from analysis of the Clostridium botulinum strain A neurotoxin gene cluster. *BMC Evolutionary Biology*, 9, pp.1–9.

Dumas, E. et al., 2009. Insight into the core and variant exoproteomes of Listeria monocytogenes species by comparative subproteomic analysis. *Proteomics*, 9(11), pp.3136–55.

Dupuy, B. & Matamouros, S., 2006. Regulation of toxin and bacteriocin synthesis in Clostridium species by a new subgroup of RNA polymerase sigma-factors. *Research in microbiology*, 157(3), pp.201–5.

Eleopra, R. et al., 1998. Different time courses of recovery after poisoning with botulinum neurotoxin serotypes A and E in humans. *Neuroscience letters*, 256(3), pp.135–8.

Erbguth, F.J., 2004. Historical notes on botulism, Clostridium botulinum, botulinum toxin, and the idea of the therapeutic use of the toxin. *Movement Disorders*, 19 Suppl 8, pp.S2–6.

Van Ermengem, E., 1979. Classics in infectious diseases. A new anaerobic bacillus and its relation to botulism. E. van Ermengem. Originally published as "Ueber einen neuen anaëroben Bacillus und seine Beziehungen zum Botulismus" in Zeitschrift für Hygiene und Infektionskrankheit. *Reviews of infectious diseases*, 1(4), pp.701–19.

Falush, D. & Bowden, R., 2006. Genome-wide association mapping in bacteria? *Trends in microbiology*, 14(8), pp.353–5.

Felsenstein, J., 1989. PHYLIP - Phylogeny Inference Package. *Cladistics*, 5(164-166).

Fenicia, L. & Anniballi, F., 2009. Infant botulism. *Ann Ist Super Sanita*, 45(2), pp.134–146.

Fillo, S. et al., 2011. Clostridium botulinum group I strain genotyping by 15-locus multilocus variable-number tandem-repeat analysis. *Journal of clinical microbiology*, 49(12), pp.4252–63.

Finegold, S.M. et al., 2002. Gastrointestinal Microflora Studies in Late-Onset Autism. *Clinical Infectious Diseases*, S.1, pp.S6–S14.

Finlay, B.B. & McFadden, G., 2006. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell*, 124(4), pp.767–82.

Fischer, A. & Montal, M., 2007. Single molecule detection of intermediates during botulinum neurotoxin translocation across membranes. *Proceedings of the National Academy of Sciences of the United States of America*, 104(25), pp.10447–52.

Fozo, E.M., Hemm, M.R. & Storz, G., 2008. Small toxic proteins and the antisense RNAs that repress them. *Microbiology and molecular biology reviews : MMBR*, 72(4), pp.579–89,

Fujinaga, Y. et al., 1997. The haemagglutinin of Clostridium botulinum type C progenitor toxin plays an essential role in binding of toxin to the epithelial cells of guinea pig small intestine, leading to the efficient absorption of the toxin. *Microbiology*, 143, pp.3841–3847.

Fujinaga, Y., 2010. Interaction of botulinum toxin with the epithelial barrier. *Journal of biomedicine & biotechnology*, 2010, p.974943.

Fujita, R. et al., 1995. Molecular characterization of two forms of nontoxic-nonhemagglutinin components of Clostridium botulinum type A progenitor toxins. *FEBS letters*, 376(1-2), pp.41–4.

Gaillot, O. et al., 2000. The ClpP serine protease is essential for the intracellular parasitism and virulence of Listeria monocytogenes. *Molecular microbiology*, 35(6), pp.1286–94.

Galazka, A. & Przybylska, A., 1999. Surveillance of foodborne botulism in Poland: 1960-1998. *Eurosurveillance*, 4(6), pp.69–72.

Gibert, M., Petit, L. & Raffestin, S., 2000. Clostridium perfringens Iota-Toxin Requires Activation of Both Binding and Enzymatic Components for Cytopathic Activity. *Infection and immunity*, 68(7), pp.3848–3853.

Gill, D.M., 1982. Bacterial toxins: a table of lethal amounts. *Microbiological reviews*, 46(1), pp.86–94.

Girardin, H. et al., 2002. Antimicrobial Activity of Foodborne Paenibacillus and Bacillus spp. against Clostridium botulinum. *Journal of food protection*, 65(5), pp.806-813.

Gold, W. & Salit, I., 1993. Aeromonas hydrophila Infections of Skin and Soft Tissue: Report of 11 Cases and Review. *Clinical Infectious Diseases*, 16(1), pp.69–74.

Gormley, F.J. et al., 2011. A 17-year review of foodborne outbreaks: describing the continuing decline in England and Wales (1992-2008). *Epidemiology and infection*, 139(5), pp.688–699.

Grant, K.A. et al., 2009. Report of two unlinked cases of infant botulism in the UK in October 2007. *Journal of Medical Microbiology*, (October 2007), pp.1601–1606.

Griffiths, B. & McClain, O., 1988. The role of iron in the growth and hemolysin (Streptolysin S) production in Streptococcus pyogenes. *J. Basic Microbiol.*, 28, pp.427–36.

Gross, W. & Smith, L., 1971. Experimental botulism in gallinaceous birds. *Avian Diseases*, 15, pp.716–722.

Gu, S. et al., 2012. Botulinum Neurotoxin Is Shielded by NTNHA in an Interlocked Complex. *Science*, 335(6071), pp.977–981.

Gupta, R.S. & Gao, B., 2009. Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus Clostridium sensu stricto (cluster I). *International Journal of Systematic and Evolutionary Microbiology*, pp.285–294.

Gutiérrez-Preciado, A. et al., 2009. Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. *Microbiology and molecular biology reviews: MMBR*, 73(1), pp.36–61.

Haas, B.J. et al., 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC genomics*, 13(1), p.734.

Hansen, M., Nielsen, A. & Molin, S., 2001. Changes in rRNA levels during stress invalidates results from mRNA blotting: fluorescence in situ rRNA hybridization permits renormalization for estimation of cellular. *Journal of bacteriology*, 183(16), p.4747.

Harrington, D., 1996. Bacterial collagenases and collagen-degrading enzymes and their potential role in human disease. *Infection and immunity*, 64(6), pp.1885–91.

Hatheway, C.L., 1990. Toxigenic clostridia. *Clinical Microbiology Reviews*, 3, pp.66–98.

Hauschild, A., 1989. Clostridium botulinum. In M. Doyle, ed. *Foodborne Bacterial Pathogens*. New York: Dekker, pp. 111–190.

Hauschild, A., 1993. Epidemiology of Human Foodborne Botulism. In A. Hauschild & K. Dodds, eds. *Clostridium botulinum: Ecology and control in food*. New York: Dekker.

Heap, J.T. et al., 2007. The ClosTron: A universal gene knock-out system for the genus Clostridium. *Journal of Microbiological Methods*, 70, pp.452 – 464.

Henderson, I. et al., 1996. Genetic characterisation of the botulinurn toxin complex of Clostridium botulinum strain NCTC 29 16. *Journal of Biological Chemistry*, 140, pp.151–158.

Heizer, E.M. et al., 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Molecular biology and evolution*, 23(9), pp.1670-80.

Hill, K.K. et al., 2007. Genetic diversity among botulinum neurotoxin-producing clostridial strains. *Journal of Bacteriology*, 189, pp.818–832.

Hill, K.K. et al., 2009. Recombination and insertion events involving the botulinum neurotoxin complex genes in Clostridium botulinum types A , B , E and F and Clostridium butyricum type E strains. *BMC Biology*, 7(66).

Hill, K.K. & Smith, T.J., 2013. Genetic Diversity Within Clostridium botulinum serotypes, botulinum neurotoxin gene clusters and toxin subtypes. In *Botulinum Neurotoxins*. pp. 1–20.

Hirst, T.R., 1995. Biogenesis of Cholera Toxin and Related Oligomeric Enterotoxins. In *Handbook of Natural Toxins*. pp. 124–171.

Holland, C. et al., 2010. Proteomic identification of secreted proteins of Propionibacterium acnes. *BMC microbiology*, 10, p.230.

Hughes, J. et al., 1981. Clinical features of types A and B food-borne botulism. *Annals of Internal Medicine* 95, pp.442–445.

Hutson, R.A. et al., 1996. Genetic Characterization of Clostridium botulinum Type A Containing Silent Type B Neurotoxin Gene Sequences *. *The Journal of Biological Chemistry*, 271(18), pp.10786 –10792.

Inoue, K. et al., 1996. Molecular composition of Clostridium botulinum type A progenitor toxins. *Infection and Immunity*, 64, pp.1589–1594.

Inui, K. et al., 2012. Toxic and nontoxic components of botulinum neurotoxin complex are evolved from a common ancestral zinc protein. *Biochemical and biophysical research communications*, 419(3), pp.500–4.

Jacobson, M.J. et al., 2008. Analysis of neurotoxin cluster genes in Clostridium botulinum strains producing botulinum neurotoxin serotype A subtypes. *Applied and Environmental Microbiology*, 74, pp.2778–2786.

Jacobson, M.J. et al., 2011. Purification, modeling, and analysis of botulinum neurotoxin subtype A5 (BoNT/A5) from Clostridium botulinum strain A661222. *Applied and environmental microbiology*, 77(12), pp.4217–22.

Johansson, J., Mandin, P. & Renzoni, A., 2002. An RNA Thermosensor Controls Expression of Virulence Genes in Listeria monocytogenes. *Cell*, 110, pp.551–561.

Jones, R.G. a et al., 2008. Development of improved SNAP25 endopeptidase immuno-assays for botulinum type A and E toxins. *Journal of immunological methods*, 329(1-2), pp.92-101.

Kaakoush, N.O. et al., 2010. The secretome of Campylobacter concisus. *The FEBS journal*, 277(7), pp.1606–17.

Katahira, J. et al., 1997. Molecular cloning and functional characterization of the receptor for Clostridium perfringens enterotoxin. *The Journal of cell biology*, 136(6), pp.1239–47.

Keller, A. et al., 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry*, 74(20), pp.5383–92.

Kirk, D.G. et al., 2012. Clostridium botulinum ATCC 3502 sigma factor K is involved with early stage sporulation. *Applied and environmental microbiology*, 78(13), pp.4590–6.

Kirk, D.G. et al., 2014. Evaluation of normalization reference genes for RT-qPCR analysis of spo0A and four sporulation sigma factor genes in Clostridium botulinum Group I strain ATCC 3502. *Anaerobe*, 26, pp.14–9.

Koepke, R., Sobel, J. & Arnon, S.S., 2008. Global Occurrence of Infant Botulism, 1976 –2006. *Pediatrics*, 122(1), pp.73–82.

Koriazova, L.K. & Montal, M., 2003. Translocation of botulinum neurotoxin light chain protease through the heavy chain channel. *Nature structural biology*, 10(1), pp.13–8.

Kroger, C. et al., 2012. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proceedings of the National Academy of Sciences*, 109(20), pp.1277–1286.

Kukreja, R. V & Singh, B.R., 2007. Comparative Role of Neurotoxin-Associated Proteins in the Structural Stability and Endopeptidase Activity of Botulinum Neurotoxin Complex Types A and E. *Biochemistry*, (17), pp.14316–14324.

Lamanna, C., 1959. The Most Poisonous Poison. *Science*, 130, pp.763–772.

Lang, A.E. et al., 2010. Photorhabdus luminescens toxins ADP-ribosylate actin and RhoA to force actin clustering. *Science*, 327(5969), pp.1139–42.

Lee, K. et al., 2013. Structure of a bimodular botulinum neurotoxin complex provides insights into its oral toxicity. *PLoS Pathogens*, 9(10), p.e1003690.

Leighton, G., 1923. *Botulism and Food Posioning (the Loch Maree Tragedy)*, Glasgow: Collins & Sons.

Lessa, F.C., Gould, C. V & McDonald, L.C., 2012. Current status of Clostridium difficile infection epidemiology. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 55 Suppl 2(Suppl 2), pp.S65–70.

Li, W. & Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13), pp.1658–9.

Lin, G. et al., 2010. Expression of the Clostridium botulinum A2 neurotoxin gene cluster proteins and characterization of the A2 complex. *Applied and environmental microbiology*, 76(1), pp.40–7.

Lindstrom, M. et al., 2001. Multiplex PCR Assay for Detection and Identification of Clostridium botulinum Types A , B , E , and F in Food and Fecal Material. *Applied and Environmental Microbiology*, 67(12), pp.5694–5699.

Lindstrom, M. & Korkeala, H., 2006. Laboratory Diagnostics of Botulism. *Clinical Microbiology Reviews*, 19(2), pp.298–314.

Lindstrom, M. et al., 2010. Clostridium botulinum in cattle and dairy products. *Critical Reviews in Food Science and Nutrition*, 50, pp.281–304.

Locht, C., Coutte, L. & Mielcarek, N., 2011. The ins and outs of pertussis toxin. *The FEBS journal*, 278(23), pp.4668–82.

Loh, E. et al., 2013. Temperature triggers immune evasion by Neisseria meningitidis. *Nature*, 502(7470), pp.237–40.

Long, S. et al., 1985. Clinical, laboratory and environmental features of infant botulism in Southeastern Pennsylvania. *Pedriatrics*, 75(5), pp.935–41.

Los, F. et al., 2013. Role of Pore-Forming Toxins in Bacterial Infectious Diseases. *Microbiol. Mol. Biol. Rev*, 77, p.173.

Lovenklev, M. et al., 2004. Relative Neurotoxin Gene Expression in Clostridium botulinum Type B , Determined Using Quantitative Reverse Transcription-PCR. *Applied and environmental microbiology*, 70(5), pp.2919–2927

Luquez, C. et al., 2005. Distribution of Botulinum Toxin-Producing Clostridia in Soils of Argentina. *Applied and environmental microbiology*, 71(7), pp.4137–4139.

Luquez, C. et al., 2012. Genetic diversity among Clostridium botulinum strains harboring bont/A2 and bont/A3 genes. *Applied and environmental microbiology*, 78(October), pp.8712–8., 2012].

De Maagd, R. a et al., 2003. Structure, diversity, and evolution of protein toxins from spore-forming entomopathogenic bacteria. *Annual review of genetics*, 37, pp.409–33.

Macdonald, T.E. et al., 2008. Differentiation of Clostridium botulinum Serotype A Strains by multi-locus vntr analysis. *Applied and Environmental Microbiology*, 74(3), pp.875–882.

Maier, T., Güell, M. & Serrano, L., 2009. Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24), pp.3966–73.

Maksymowych, A.B. & Simpson, L.L., 1998. Binding and Transcytosis of Botulinum Neurotoxin by Polarized Human Colon Carcinoma Cells. *The Journal of Biological Chemistry*, 273(34), pp.21950 –21957.

Marvaud, J.C. et al., 1998. botR / A is a positive regulator of botulinum neurotoxin and associated non-toxin protein genes in Clostridium botulinum A. *Molecular Microbiology*, 29, pp.1009–1018.

Matsumura, T. et al., 2008. The HA proteins of botulinum toxin disrupt intestinal epithelial intercellular junctions to increase toxin absorption. *Cell. Microbiol*, 10, pp.355–364.

Matsushita, O. & Okabe, a, 2001. Clostridial hydrolytic enzymes degrading extracellular components. *Toxicon : official journal of the International Society on Toxinology*, 39(11), pp.1769–80.

Mclauchlin, J., Grant, K.A. & Little, C.L., 2006. Food-borne botulism in the United Kingdom. *Journal of Public Health*, 28(4), pp.337–342.

Mclauchlin, J. & Grant, K.A., 2007. Clostridium botulinum and Clostridium perfringens. In S. Simjee, ed. *Foodborne Diseases*. Totowa: Humana Press, pp. 41–78.

Midura, T. & Arnon, S.S., 1976. Infant botulism. Identification of Clostridium botulinum and its toxins in faeces. *Lancet*, 2(7992), pp.934–6.

Mildenhall, D.C., Wiltshire, P.E.J. & Bryant, V.M., 2006. Forensic palynology: why do it and how it works. *Forensic science international*, 163(3), pp.163–72.

Minami, J. et al., 1997. Lambda-toxin of Clostridium perfringens activates the precursor of epsilon-toxin by releasing its N- and C-terminal peptides. *Microbiology and immunology*, 41(7), pp.527–35.

Monod, J., 1949. The growth of bacterial cultures. *Annual review of microbiology*, 3, pp.371–394.

Montal, M., 2010. Botulinum neurotoxin: a marvel of protein design. *Annual review of biochemistry*, 79, pp.591–617.

Montecucco, C. & Schiavo, G., 1994. Mechanism of action of tetanus and botulinum neurotoxins. *Journal of Biological Chemistry*, 13, pp.1 – 8.

Montecucco, C., Papini, E. & Schiavo, G., 1994. Bacterial protein toxins penetrate cells via a four-step mechanism. *FEBS letters*, 346(1), pp.92–8.

Munro, N.B., Ambrose, K.R. & Watson, A.P., 1994. Toxicity of the Organophosphate Chemical Warfare Agents GA, GB and VX: Implications for Public Protection. *Environmental Health Perspectives*, 102(1), pp.18–38.

Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R., 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)*, 25(10), pp.1335–7.

Nesvizhskii, A.I. et al., 2003. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry abilities that proteins are present in a sample on the basis. *Anal. Chem.*, 75(17), pp.4646–4658.

Nicholson, J.K. et al., 2012. Host-gut microbiota metabolic interactions. *Science (New York, N.Y.)*, 336(6086), pp.1262–7.

Nishikawa, A. et al., 2004. The receptor and transporter for internalization of Clostridium botulinum type C progenitor toxin into HT-29 cells. *Biochemical and biophysical research communications*, 319(2), pp.327–33.

Niwa, K. et al., 2007. Role of nontoxic components of serotype D botulinum toxin complex in permeation through a Caco-2 cell monolayer, a model for intestinal epithelium. *FEMS Microbiology Letters*, 49, pp.346–352.

Ohishi, I., Sugii, S. & Sakaguchi, G., 1977. Oral Toxicities of Clostridium botulinum Toxins in Response to Molecular Size. *Infection and Immunity*, 16(1), p.107.

Olsen, S. J. et al. "Surveillance for foodborne-disease outbreaks--United States, 1993-1997." MMWR CDC Surveill Summ. 49.1 (2000): 1-62.

Oshlack, A., Robinson, M.D. & Young, M.D., 2010. From RNA-seq reads to differential expression results. *Genome biology*, 11(12), p.220.

Papenfort, K. & Vogel, J., 2010. Regulatory RNA in bacterial pathogens. *Cell host & microbe*, 8(1), pp.116–27. *f*Barash

Peck, M.W., 2009. Biology and Genomic Analysis of Clostridium botulinum. In *Advances in Microbial Physiology*. pp. 183 – 265.

Peck, M.W., Stringer, S.C. & Carter, A.T., 2011. Clostridium botulinum in the post-genomic era. *Food microbiology*, 28(2), pp.183–91.

Pellizzari, R. et al., 1999. Tetanus and botulinum neurotoxins: mechanism of action and therapeutic uses. *Phil. Trans. R. Soc. Lond. B*, 354, pp.259–268.

Pfaffl, M.W., 2006. Relative quantification. In T. Dorak, ed. *Real-time PCR*. Taylor & Francis, pp. 63–82.

Pickett, J. et al., 1976. Syndrome of Botulism in Infancy: Clinical and Electrophysiologic Study. *New England Journal of Medicine*, 295, pp.770–772.

Pocsfalvi, G. et al., 2008. Proteomic analysis of exoproteins expressed by enterotoxigenic Staphylococcus aureus strains. *Proteomics*, 8(12), pp.2462–76.

Popoff, M.R., 1995. Ecology of Neurotoxigenic Strains of Clostridia. In C. Montecucco, ed. *Clostridial Neurotoxins*. Springer, pp. 1–30.

Porcelli, I., Reuter, M. & Pearson, B., 2013. Parallel evolution of genome structure and transcriptional landscape in the Epsilonproteobacteria. *BMC genomics*, 14, p.616.

Pott, J. et al., 2012. Age-dependent TLR3 expression of the intestinal epithelium contributes to rotavirus susceptibility. *PLoS pathogens*, 8(5), p.e1002670.

Puerta-Fernandez, E. et al., 2006. Identification of a large noncoding RNA in extremophilic eubacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51), pp.19490–5.

Quevillon, E. et al., 2005. InterProScan: protein domains identifier. *Nucleic acids research*, 33(Web Server issue), pp.W116–20.

Raffestin, S. et al., 2005. BotR / A and TetR are alternative RNA polymerase sigma factors controlling the expression of the neurotoxin and associated protein genes in Clostridium botulinum type A and Clostridium tetani. *Molecular Microbiology*, 55, pp.235–249.

Rao, S. et al., 2007. Variations in expression and release of botulinum neurotoxin in Clostridium botulinum type A strains. *Foodborne pathogens and disease*, 4, pp.201–7.

Robinson, M.D. & Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3), p.R25.

Rodríguez Jovita, M., Collins, M.D. & East, a K., 1998. Gene organization and sequence determination of the two botulinum neurotoxin gene clusters in Clostridium botulinum type A(B) strain NCTC 2916. *Current microbiology*, 36(4), pp.226–31.

Rossetto, O. et al., 1994. SNARE motif and neurotoxins. *Nature*, 372, pp.415–16.

Rumbaugh, K. & Griswold, J., 1999. Contribution of Quorum Sensing to the Virulence ofPseudomonas aeruginosa in Burn Wound Infections. *Infection and immunity*, 67(11), p.5854.

Rutherford, K. et al., 2000. Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)*, 16(10), pp.944–5.

Sakaguchi, G., 1982. Clostridium botulinum toxins. *Pharmacology therapeutics*, 19(2), pp.165–194.

Schantz, E.J. & Johnson, E.A., 1992. Properties and use of botulinum toxin and other microbial neurotoxins in medicine. *Microbiological Reviews*, 56, p.80.

Schiavo, G. et al., 1992. Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature*, 359, pp.832–35.

Schramm, G., Bruchhaus, I. & Roeder, T., 2000. A simple and reliable 5'-RACE approach. *Nucleic acids research*, 28(22), p.E96.

Schwarz, K. et al., 2007. A Standard Operating Procedure (SOP) for the preparation of intra- and extracellular proteins of Clostridium acetobutylicum for proteome analysis. *Journal of Microbiological Methods*, 68, pp.396 – 402.

Sebaihia, M. et al., 2007. Genome sequence of a proteolytic (Group I) Clostridium botulinum strain Hall A and comparative analysis of the clostridial genomes. *Genome Research*, 17, pp.1082–1092.

Sergeant, M. et al., 2003. Interactions of Insecticidal Toxin Gene Products from Xenorhabdus nematophilus PMFI296. *Applied and Environmental Microbiology*, 69(6), pp.3344–3349.

Shabbiri, K. et al., 2013. Charting the cellular and extracellular proteome analysis of Brevibacterium linens DSM 20158 with unsequenced genome by mass spectrometry-driven sequence similarity searches. *Journal of Proteomics*, 83(i), pp.99–118.

Sharma, C.M. et al., 2010. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, 464(7286), pp.250–5.

Sharma, C. & Heidrich, N., 2012. Small RNAs and virulence in bacterial pathogens. *RNA biology*, 9(4), pp.361–363.

Shapiro, R.L., Hatheway, C. & Swerdlow, D.L., 1998. Botulism in the United States: A Clinical and Epidemiologic Review. , 129(3), pp.221–228.

Sharma, D., 1999. *Toxin production by Clostridium botulinum*. University of East Anglia.

Sheets, J.J. et al., 2011. Insecticidal toxin complex proteins from Xenorhabdus nematophilus: structure and pore formation. *The Journal of Biological Chemistry*, 286(26), pp.22742–9.

Sheppard, Y.D. et al., 2012. Intestinal Toxemia Botulism in 3 Adults, Ontario, Canada, 2006–2008. *Emerging Infectious Diseases*, 18(1), pp.2006–2008.

Sheth, A.N. et al., 2008. International outbreak of severe botulism with prolonged toxemia caused by commercial carrot juice. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 47(10), pp.1245–51.

Shin, N. et al., 2006. Determination of Neurotoxin Gene Expression in Clostridium botulinum Type A by Quantitative RT-PCR. *Molecules and Cells*, 22(3), pp.336–342.

Shimizu, T., Shima, K. & Yoshino, K., 2002. Proteome and transcriptome analysis of the virulence genes regulated by the VirR/VirS system in Clostridium perfringens. *Journal of Bacteriology* 184(10), pp.2587–2594.

Singh, B.R., Li, B. & Read, D., 1995. Botulinum versus tetanus neurotoxins: why is botulinum neurotoxin but not tetanus neurotoxin a food poison? *Toxicon*, 33(12), pp.1541–1547.

Singh, A.K. et al., 2013. Purification and characterization of neurotoxin complex from a dual toxin gene containing Clostridium Botulinum Strain PS-5. *The protein journal*, 32(4), pp.288–96.

Simpson, L.L., 2004. Identification of the Major Steps in Botulinum Toxin Action. *Annu. Rev. Pharmacol. Toxicol*, 44, pp.167–93.

Sivaraman, T. et al., 1997. The mechanism of trichloroacetic acid induced protein precipitation. *J. Protein Chem*, 16(4), pp.291–7.

Skarin, H. & Segerman, B., 2011. Horizontal gene transfer of toxin genes in Clostridium botulinum. *Mobile Genetic Elements*, 1(3), pp.213–215.

Smith, L., 1978. The occurrence of Clostridium botulinum and Clostridium tetani in the soil of the United States. *Health Laboratory Science*, 15(2), pp.74–80.

Smith, G.R. & Young, a M., 1980. Clostridium botulinum in British soil. *The Journal of hygiene*, 85(2), pp.271–4.

Smith, T.J. et al., 2007. Analysis of the neurotoxin complex genes in Clostridium botulinum A1-A4 and B1 strains: BoNT/A3, /Ba4 and /B1 clusters are located within plasmids. *PLoS ONE*, 12, p.e1271.

Smith, D. & Chapman, M., 2010. Economical evolution: microbes reduce the synthetic cost of extracellular proteins. *MBio*, 1(3).

Sobel, J. et al., 2004. Foodborne Botulism in the United States, 1990–2000. *Emerging Infectious Diseases*, 10(9), p.1606.

Sobel, J., 2005. Botulism. *Clinical Infectious Diseases*, 41, pp.1167–1173.

Song, J., Gao, X. & Galán, J.E., 2013. Structure and function of the Salmonella Typhi chimaeric A(2)B(5) typhoid toxin. *Nature*, 499(7458), pp.350–4.

Sonnabend, O. et al., 1981. Isolation of Clostridium botulinum Type G and Identification of Type G Botulinal Toxin in Humans : Report of Five Sudden Unexpected Deaths. *Journal of Infectious Diseases*, 143(1), pp.22–27.

Sonnabend, O. et al., 1985. Continuous microbiological and pathological study of 70 sudden and unexpected infant deaths: toxigenic intestinal Clostridium botulinum infection in 9 cases of sudden infant death syndrome. *The Lancet*, 325(8243), pp.237–241.

Stebbins, C.E., 2013. Bacteriology: toxins in tandem. *Nature*, 499(7458), p.293.

Stein, P.E. et al., 1994. The crystal structure of pertussis toxin. *Structure*, 2(1), pp.45–57.

Stenmark, P. et al., 2008. Crystal structure of botulinum neurotoxin type A in complex with the cell surface co-receptor GT1b-insight into the toxin-neuron interaction. C. E. Stebbins, ed. *PLoS Pathogens*, 4(8), p.e1000129.

Suen, J.C. et al., 1988. Clostridium argentinense sp. nov.: A Genetically Homogeneous Group Composed of All Strains of Clostridium botulinum Toxin Type G and Some Nontoxigenic Strains Previously Identified as Clostridium subterminale or Clostridium hastiforme. *International Journal of Systematic Bacteriology*, 38(4), pp.375–381.

Sugawara, Y. et al., 2010. Botulinum hemagglutinin disrupts the intercellular epithelial barrier by directly binding E-cadherin. *The Journal of cell biology*, 189(4).

Sugii, S. & Sakaguchi, G., 1975. Molecular construction of Clostridium botulinum type A toxins. *Infect. Immun.*, 12, pp.1262–1270.

Tamura, M. et al., 1982. Subunit structure of islet-activating protein, pertussis toxin, in conformity with the A-B model. *Biochemistry*, 21(22), pp.5516–22.

Tamura, K. et al., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10), pp.2731–9.

Taylor, S.M. et al., 2010. Wound botulism complicating internal fixation of a complex radial fracture. *Journal of clinical microbiology*, 48(2), pp.650–3.

Telford, J.L. et al., 2006. Pili in Gram-positive pathogens. *Nature Reviews Microbiology*, 4, pp.509–519.

Tepp, W.H., Lin, G. & Johnson, E. a, 2012. Purification and Characterization of a Novel Botulinum Neurotoxin Subtype A3. *Applied and environmental microbiology*, 78(9), pp.3108–13.

Terilli, R.R. et al., 2011. A historical and proteomic analysis of botulinum neurotoxin type/G. *BMC microbiology*, 11(1), p.232.

Terra, W.R., 1996. Biology of the Insect Midgut. In M. Lehane & P. Billingsley, eds. Springer, pp. 153–157.

Thomas, M.K. et al., 2013. Estimates of the Burden of Foodborne Illness in Canada for 30 Specified Pathogens and Unspecified Agents, Circa 2006. *Foodborne pathogens and disease*, 10(7), pp.639–48.

Thompson, L.J. et al., 2003. Gene expression profiling of Helicobacter pylori reveals a growth-phase-dependent switch in virulence gene expression. *Infection and ...*, 71(5), pp.2643–2655.

Toledo-Arana, A., Repoila, F. & Cossart, P., 2007. Small noncoding RNAs controlling pathogenesis. *Current opinion in microbiology*, 10(2), pp.182–8.

Tsai, Y.C. et al., 2010. Targeting botulinum neurotoxin persistence by the ubiquitin-proteasome system. *PNAS*, 107(38), pp.16554–59.

Tsukamoto, K. et al., 2005. Binding of Clostridium botulinum type C and D neurotoxins to ganglioside and phospholipid; novel insights into the receptor for clostridial neurotoxins. *The Journal of biological chemistry*, 280(42), pp.35164–71.

Tversky, A. & Kahneman, D., 1971. Belief in the law of small numbers. *Psychological Bulletin*, 76(2), pp.105–110.

Underwood, K. et al., 2007. Infant botulism: a 30-year experience spanning the introduction of botulism immune globulin intravenous in the intensive care unit at Childrens Hospital Los Angeles. *Pediatrics*, 120(6), pp.e1380–5.

Verderio, C. et al., 2006. Entering neurons: botulinum toxins and synaptic vesicle recycling. *EMBO reports*, 7(10), pp.995–9.

Vos, P. et al., 2009. *Bergey's Manual of Systematic Bacteriology: The Firmicutes*, Springer.

Wagner, V., Bushnell, D. & Passador, L., 2003. Microarray Analysis of Pseudomonas aeruginosa Quorum-Sensing Regulons: Effects of Growth Phase and Environment. *Journal of bacteriology*, 185(7), p.2080.

Wallace, J.G., Zhou, Z. & Breaker, R.R., 2012. OLE RNA protects extremophilic bacteria from alcohol toxicity. *Nucleic acids research*, (19), pp.1–10.

Walz, A. et al., 2007. Bacillus anthracis secretome time course under host-simulated conditions and identification of immunogenic proteins. *Proteome science*, 5, 11.

Wang, Y. et al., 2012. Genome-wide dynamic transcriptional profiling in Clostridium beijerinckii NCIMB 8052 using single-nucleotide resolution RNA-Seq. *BMC genomics*, 13(1), p.102.

Waterfield, N.R. et al., 2001. The tc genes of Photorhabdus: a growing family. *Trends in microbiology*, 9(4), pp.185–91.

Waters, L.S. & Storz, G., 2009. Regulatory RNAs in bacteria. *Cell*, 136(4), pp.615–28.

Weber, J.T. et al., 1993. Wound botulism in a patient with a tooth abscess: case report and review. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 16(5), pp.635–9.

Van Wely, K.H. et al., 2001. Translocation of proteins across the cell envelope of Gram-positive bacteria. *FEMS microbiology reviews*, 25(4), pp.437–54.

Wheeler, C. et al., 2009. Sensitivity of Mouse Bioassay in Clinical Wound Botulism. *Clinical Infectious Diseases*, 6403, pp.1669–1673.

Whitemarsh, R.C.M. et al., 2013. Characterization of botulinum neurotoxin a subtypes 1 through 5 by investigation of activities in mice, in neuronal cell cultures, and in vitro. *Infection and immunity*, 81(10), pp.3894–902.

Woodruff, B. a et al., 1992. Clinical and laboratory comparison of botulism from toxin types A, B, and E in the United States, 1975-1988. *The Journal of infectious diseases*, 166(6), pp.1281–6.

Wylie, C.E. & Proudman, C.J., 2009. Equine Grass Sickness: Epidemiology, Diagnosis, and Global Distribution. *Veterinary Clinics of North America Equine Practices*, 25, pp.381–99.

Yu, C. & Chen, Y., 2006. Prediction of protein subcellular localization. *Proteins: Structure, Function and Bioinformtatics* 651(December 2005), pp.643–651.

Yu, N.Y. et al., 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics (Oxford, England)*, 26(13), pp.1608–15.

Zhang, Z. et al., 2013. Two-component signal transduction system CBO0787/CBO0786 represses transcription from botulinum neurotoxin promoters in Clostridium botulinum ATCC 3502. *PLoS pathogens*, 9(3), p.e1003252.

Zhou, C.E. et al., 2007. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic acids research*, 35(Database issue), pp.D391–4

Zhou, M. et al., 2008. LocateP: genome-scale subcellular-location predictor for bacterial proteins. *BMC bioinformatics*, 9, p.173.

1.1. Appendix

Appendix Table 1: HMMer matches for P-47 family proteins from non-C. botulinum organisms.

Ricketsiella grylli

Query	Match uniprot	Match protein name	Match protein species	Match score
>Annotation was generated automatically without man - 717439: 719067 MW: 59884.02	A8PMG2_9COXI	Putative uncharacterized protein (gene: RICGR_0717)	Rickettsiella grylli	0.00E+00
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	5.10E-201
	B2VCQ8_ERWT9	Putative uncharacterized protein (gene: ETA_30270)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	3.80E-169
	E9N6Q5_CLOBO	OrfX2 (gene: orfX2)	Clostridium botulinum	3.70E-32
	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	2.40E-31
	F7SP90_9GAMM	Putative uncharacterized protein (gene: GME_11572)	Halomonas sp. TD01	3.20E-18
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	2.30E-15
	B1QQF2_CLOBO	Toxin complex component ORF-X3 (gene: CBB_A0177)	Clostridium botulinum Bf	1.80E-08
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	7.00E-08
	G5C996_HETGA	Obscurin (gene: GW7_06474)	Heterocephalus glaber	3.40E-05
	F7CQP1_MACMU	Uncharacterized protein (gene: OBSCN)	Macaca mulatta	3.80E-05
	F7C4M0_CALJA	Uncharacterized protein (gene: OBSCN)	Callithrix jacchus	5.60E-05
	F1LVE2_RAT	Uncharacterized protein (Fragment)	Rattus norvegicus	6.30E-05

	H2N3G8_PONAB	Uncharacterized protein (gene: OBSCN)	Pongo abelii	6.60E-05
>Annotation was generated automatically without man - 719085: 720371 MW: 47997.836	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	7.90E-292
	D2TXI9_9ENTR	Putative uncharacterized protein (gene: ARN_081010)	Arsenophonus nasoniae	4.80E-154
	B2VCQ7_ERWT9	Putative uncharacterized protein (gene: ETA_30260)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	4.00E-95
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	1.50E-31
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	1.50E-20
	Q6RI05_CLOBO	P-47 (gene: p47)	Clostridium botulinum	2.30E-17
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	2.00E-10
	A7GBF7_CLOBL	Toxin complex component ORF-X3 (gene: CLI_0845)	Clostridium botulinum (strain Langeland / NCTC 10281 / Type F)	8.50E-10
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	1.10E-07
	F7SWZ6_ALCXX	Putative uncharacterized protein (gene: AXXA_06258)	Achromobacter xylosoxidans AXX-A	1.30E-07
	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	9.50E-06
>Annotation was generated automatically without man - 720419: 720865 MW: 17117.98	A8PMG4_9COXI	Putative uncharacterized protein (gene: RICGR_0719)	Rickettsiella grylli	6.40E-94
	D2TXJ1_9ENTR	Putative uncharacterized protein (gene: ARN_08130)	Arsenophonus nasoniae	1.30E-30

	B2VCQ5_ERWT9	Putative uncharacterized protein (gene: ETA_30250)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	8.00E-12
	E5B9Q2_ERWAM	Putative uncharacterized protein (gene: EAIL5_3309)	Erwinia amylovora ATCC BAA-2158	3.20E-11
>Annotation was generated automatically without man - 844573: 845499 MW: 34598.996	A8PMG6_9COXI	Shiga toxin A-chain (RRNA N-glycosidase) (gene: RICGR_0720)	Rickettsiella grylli	1.90E-200
	Q2HWU1_ECOLX	Shiga toxin 2 variant f A-subunit (gene: stx2fA)	Escherichia coli O63:HNM	4.50E-60
	B3GK88_ECOLX	Shiga toxin 1 subunit A (gene: stxA1)	Escherichia coli	3.70E-58
	Q2L9B4_ACIHA	Shiga toxin II subunit A (gene: stx2A)	Acinetobacter haemolyticus	3.50E-56
	D5WND3_BURSC	Ribosome-inactivating protein (Precursor) (gene: BC1002_7046)	Burkholderia sp. (strain CCGE1002)	6.00E-32
	D6MWK5_RICCO	rRNA N-glycosidase (Fragment)	Ricinus communis	9.30E-06
>Annotation was generated automatically without man - 844116: 844526 MW: 15174.371	A8PMG7_9COXI	Putative uncharacterized protein (gene: RICGR_0721)	Rickettsiella grylli	2.60E-89
	C4SAT8_YERMO	Putative uncharacterized protein (gene: ymoll0001_30930)	Yersinia mollaretii ATCC 43969	6.90E-08
	B6SD00_9VIRU	Putative uncharacterized protein P10 (gene: P10)	Bacteriophage APSE-4	1.50E-05
	B6SD14_9VIRU	Putative uncharacterized protein X (gene: X)	Bacteriophage APSE-7	1.80E-05

P. larvae

Query	Match uniprot	Match protein name	Match protein species	Match score
>PlarIB_020100022898 - 60616: 61908 MW: 48777.277	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	7.60E-147
	E9N6Q5_CLOBO	OrfX2 (gene: orfX2)	Clostridium botulinum	1.50E-76
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	1.00E-18

	A8PMG2_9COXI	Putative uncharacterized protein (gene: RICGR_0717)	Rickettsiella grylli	8.90E-16
	F7SWZ6_ALCXX	Putative uncharacterized protein (gene: AXXA_06258)	Achromobacter xylosoxidans AXX-A	3.90E-14
	F7SP90_9GAMM	Putative uncharacterized protein (gene: GME_11572)	Halomonas sp. TD01	2.90E-13
	B2VCQ8_ERWT9	Putative uncharacterized protein (gene: ETA_30270)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	9.10E-11
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	1.20E-07
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.70E-06
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	1.40E-05
	C3KS19_CLOB6	Toxin complex component ORF-X3 (gene: CLJ_0010)	Clostridium botulinum (strain 657 / Type Ba4)	3.40E-03
misc_feature complement(5378..6775)	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	5.10E-279
/note="potential frameshift: common BLAST hit:	E9N6Q4_CLOBO	OrfX3 (gene: orfX3)	Clostridium botulinum	3.00E-122
gi|153939271|ref|YP_001390117.1| toxin complex component	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	4.00E-62
ORF-X3"	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	3.40E-42
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	6.80E-37
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	2.00E-20
	B2V3V1_CLOBA	Toxin complex component ORF-X2 (gene: CLH_1114)	Clostridium botulinum (strain Alaska E43 / Type E3)	1.50E-15
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	5.10E-11

misc_feature complement(3984..5274)	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	9.00E-201
/note="potential frameshift: common BLAST hit:	B1L2G3_CLOBM	P-47 protein (gene: p47)	Clostridium botulinum (strain Loch Maree / Type A3)	3.40E-52
gi|226948034|ref|YP_002803125.1| P-47 protein"	Q3SU90_NITWN	Putative uncharacterized protein (gene: Nwi_0887)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	1.90E-28
	D2TXI9_9ENTR	Putative uncharacterized protein (gene: ARN_081010)	Arsenophonus nasoniae	1.40E-18
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	1.20E-14
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	3.00E-14
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	1.70E-11
	F7SP89_9GAMM	P-47 protein (gene: GME_11567)	Halomonas sp. TD01	1.70E-07
	E9N6Q4_CLOBO	OrfX3 (gene: orfX3)	Clostridium botulinum	3.60E-05
>PlarIB_02010002287 3 - 64828: 65511 MW: 25735.86	H3SHP6_9BACL	Protective antigen (gene: PDENDC454_15317)	Paenibacillus dendritiformis C454	0.00E+00
and	F7J0A3_CLOPE	Iota toxin component Ib (gene: ibp)	Clostridium perfringens	1.10E-80
misc_feature complement(1378..3297)	O06498_9FIRM	Sb component (gene: sbs)	Clostridium spiroforme	1.30E-78
/note="potential frameshift: common BLAST hit:	C6DXY3_CLOBO	C2 toxin, component II (gene: CLG_0130)	Clostridium botulinum D str. 1873	6.80E-71
gi|260687838|ref|YP_003218972.1| ADP-ribosyltransferase	B3ZYH5_BACCE	Iota toxin component Ib (gene: BC03BB108_C0181)	Bacillus cereus 03BB108	1.70E-62
binding protein"	C3FB27_BACTU	Iota toxin component Ib (gene: bthur0007_55280)	Bacillus thuringiensis serovar monterrey BGSC 4AJ1	9.60E-61
	F7TW88_BRELA	Protective antigen (gene: pagA2)	Brevibacillus laterosporus LMG 15441	1.50E-58

	C3LLD4_BACAC	Protective antigen (gene: pagA)	Bacillus anthracis (strain CDC 684 / NRRL 3495)	1.50E-49
	G9XXH2_SPIME	Protective antigen (gene: SPM_01175)	Spiroplasma melliferum KC3	2.20E-45

Erwinia tasmaniensis

Query	Match uniprot	Match protein name	Match protein species	Match score
>silverDB:etchr03014 - 503464: 505140 MW: 60636.168	B2VCQ8_ERWT9	Putative uncharacterized protein (gene: ETA_30270)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	0.00E+00
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	1.20E-169
	A8PMG2_9COXI	Putative uncharacterized protein (gene: RICGR_0717)	Rickettsiella grylli	8.90E-169
	E9N6Q5_CLOBO	OrfX2 (gene: orfX2)	Clostridium botulinum	5.40E-30
	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	5.70E-28
	B0FNR1_CLOBO	OrfX3 (Fragment)	Clostridium botulinum	5.60E-15
	C5UY10_CLOBO	Toxin complex component ORF-X3 (gene: CLO_2647)	Clostridium botulinum E1 str. 'BoNT E Beluga'	7.40E-07
	F7SWZ6_ALCXX	Putative uncharacterized protein (gene: AXXA_06258)	Achromobacter xylosoxidans AXX-A	4.60E-06
>silverDB:etchr03013 - 505151: 506356 MW: 44426.043	B2VCQ7_ERWT9	Putative uncharacterized protein (gene: ETA_30260)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	2.20E-274
	D2TXI9_9ENTR	Putative uncharacterized protein (gene: ARN_081010)	Arsenophonus nasoniae	2.80E-119
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	2.50E-97
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	5.40E-20
	Q6RI05_CLOBO	P-47 (gene: p47)	Clostridium botulinum	3.40E-15

	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.50E-13
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	4.10E-10
	C9WWY1_CLOBO	ORFX3 (Fragment)	Clostridium botulinum	1.70E-09
>silverDB:etchr03012 - 506359: 506814 MW: 17165.527	B2VCQ5_ERWT9	Putative uncharacterized protein (gene: ETA_30250)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	6.20E-94
	A8PMG4_9COXI	Putative uncharacterized protein (gene: RICGR_0719)	Rickettsiella grylli	1.10E-11
	D2TXJ1_9ENTR	Putative uncharacterized protein (gene: ARN_08130)	Arsenophonus nasoniae	9.60E-06
>silverDB:etchr03011 - 506899: 508029 MW: 42732.51	B2VCQ4_ERWT9	Putative uncharacterized protein (gene: ETA_30240)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	1.30E-259
	B6SD26_9VIRU	Putative uncharacterized protein Y (gene: Y)	Bacteriophage APSE-3	5.00E-67
	B8NDI1_ASPFN	Putative uncharacterized protein (gene: AFLA_059200)	Aspergillus flavus (strain ATCC 200026 / FGSC A1120 / NRRL 3357 / JCM 12722 / SRRC 167)	6.70E-10
	A9ZJ22_COXBE	Putative uncharacterized protein (gene: COXBURSA334_A0135)	Coxiella burnetii RSA 334	3.70E-04
>silverDB:etchr03010 - 508094: 513058 MW: 181298.03	B2VCQ3_ERWT9	Nematicidal protein (gene: rhsA)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	0.00E+00
	B6SD27_9VIRU	Putative YD-repeat toxin (gene: Z)	Bacteriophage APSE-3	0.00E+00
	C4T5C2_YERIN	Rhs family protein (gene: yinte0001_39190)	Yersinia intermedia ATCC 29909	0.00E+00
	H1ZAP4_9FLAO	RHS repeat-associated core domain-containing protein (gene: Myrod_3544)	Myroides odoratus DSM 2801	1.90E-284
	H3RHE1_ERWST	Putative uncharacterized protein (gene: CKS_5332)	Pantoea stewartii subsp. stewartii DC283	5.20E-260

	Q9EVR7_XENBV	Nematicidal protein 2 (gene: xnp2)	Xenorhabdus bovienii	1.90E-252
	B2Q268_PROST	Putative uncharacterized protein (gene: PROSTU_02624)	Providencia stuartii ATCC 25827	1.50E-246
	G7XWY4_ASPKW	RHS repeat protein (gene: AKAW_09557)	Aspergillus kawachii (strain NBRC 4308)	4.60E-188
	G3JUS5_CORMM	RHS Repeat protein (gene: CCM_09498)	Cordyceps militaris (strain CM01)	2.30E-175
	F3DCZ3_9PSED	YD repeat-containing protein (gene: PSYAE_09539)	Pseudomonas syringae pv. aesculi str. 0893_23	3.00E-146
	Q7N4A7_PHOLL	Putative uncharacterized protein plu2442 (gene: plu2442)	Photorhabdus luminescens subsp. laumondii (strain TT01)	5.10E-120
	O52883_COXBE	Putative uncharacterized protein orf 526 (gene: orf 526)	Coxiella burnetii	1.90E-93
	A7MZ70_VIBHB	Putative uncharacterized protein (gene: VIBHAR_01748)	Vibrio harveyi (strain ATCC BAA-1116 / BB120)	4.90E-86
	C7HJS3_CLOTM	YD repeat protein (Precursor) (gene: ClothDRAFT_2932)	Clostridium thermocellum DSM 2360	6.70E-12

Arsenophonus nasoniae

Query	Match uniprot	Match protein name	Match protein species	Match score
>ARN_08100 - 4071: 5693 MW: 59243.086	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	0.00E+00
	A8PMG2_9COXI	Putative uncharacterized protein (gene: RICGR_0717)	Rickettsiella grylli	1.20E-200
	B2VCQ8_ERWT9	Putative uncharacterized protein (gene: ETA_30270)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	1.20E-169
	C5UY11_CLOBO	Toxin complex component ORF-X2 (gene: CLO_2648)	Clostridium botulinum E1 str. 'BoNT E Beluga'	1.80E-30
	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	4.00E-29
	C9WWY1_CLOBO	ORFX3 (Fragment)	Clostridium botulinum	8.00E-16

	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	3.40E-17
>ARN_081010 - 5721: 6992 MW: 47093.24	D2TXI9_9ENTR	Putative uncharacterized protein (gene: ARN_081010)	Arsenophonus nasoniae	1.40E-285
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	2.40E-154
	B2VCQ7_ERWT9	Putative uncharacterized protein (gene: ETA_30260)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	2.10E-117
	P71115_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	2.00E-13
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.20E-12
	B0FNR1_CLOBO	OrfX3 (Fragment)	Clostridium botulinum	3.50E-10
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	3.80E-07
	C4IHL6_CLOBU	Toxin complex component ORF-X2 (gene: CLP_2747)	Clostridium butyricum E4 str. BoNT E BL5262	3.20E-04
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	3.90E-03
>ARN_08120 - 7022: 8646 MW: 59423.383	exact match isnt in uniprot			
	E3DHB7_ERWSE	Putative toxin-like protein (gene: EJP617_07250)	Erwinia sp. (strain Ejp617)	4.70E-62
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	1.80E-60
	A8PMG2_9COXI	Putative uncharacterized protein (gene: RICGR_0717)	Rickettsiella grylli	4.60E-58
	B2V3V1_CLOBA	Toxin complex component ORF-X2 (gene: CLH_1114)	Clostridium botulinum (strain Alaska E43 / Type E3)	3.40E-26
	C4IHL5_CLOBU	Toxin complex component ORF-X3 (gene: CLP_2746)	Clostridium butyricum E4 str. BoNT E BL5262	1.40E-09
	P71113_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	5.70E-03

>ARN_08130 - 8728: 9174 MW: 16951.338	D2TXJ1_9ENTR	Putative uncharacterized protein (gene: ARN_08130)	Arsenophonus nasoniae	4.30E-90
	A8PMG4_9COXI	Putative uncharacterized protein (gene: RICGR_0719)	Rickettsiella grylli	4.50E-31
	B2VCQ5_ERWT9	Putative uncharacterized protein (gene: ETA_30250)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	2.70E-06
>ARN_08140 - 9255: 12254 MW: 114877.19	D2TXJ2_9ENTR	Nematicidal protein (gene: ARN_08140)	Arsenophonus nasoniae	0.00E+00
	Q7N4A7_PHOLL	Putative uncharacterized protein plu2442 (gene: plu2442)	Photorhabdus luminescens subsp. laumondii (strain TT01)	5.40E-173
	H3RHE1_ERWST	Putative uncharacterized protein (gene: CKS_5332)	Pantoea stewartii subsp. stewartii DC283	7.40E-136
	B2VJI1_ERWT9	Similar to Nematicidal protein 2 (gene: xnp2)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	4.80E-121
	G4KI80_YEREN	Rhs family protein (gene: IOK_17811)	Yersinia enterocolitica subsp. palearctica PhRBD_Ye1	1.80E-113
	Q9EVR7_XENBV	Nematicidal protein 2 (gene: xnp2)	Xenorhabdus bovienii	1.20E-104
>ARN_08150 - 12390: 14057 MW: 61276.785	exact match isnt in uniprot			
	Q7N4A7_PHOLL	Putative uncharacterized protein plu2442 (gene: plu2442)	Photorhabdus luminescens subsp. laumondii (strain TT01)	4.10E-98
	C4T5C2_YERIN	Rhs family protein (gene: yinte0001_39190)	Yersinia intermedia ATCC 29909	4.60E-82
	Q9S6J1_COXBE	Putative uncharacterized protein	Coxiella burnetii	9.20E-79
	B2VJI1_ERWT9	Similar to Nematicidal protein 2 (gene: xnp2)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	4.00E-76

	Q7MB38_PHOLL	Similar to the nematicidal protein 2. Probable membrane protein (gene: plu2222)	Photorhabdus luminescens subsp. laumondii (strain TT01)	1.90E-65

Halomonas

Query	Match uniprot	Match protein name	Match protein species	Match score
>GME_11567 - 5384: 6736 MW: 48127.82	F7SP89_9GAMM	P-47 protein (gene: GME_11567)	Halomonas sp. TD01	3.30E-302
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	2.30E-13
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	2.90E-10
	P71115_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	7.60E-09
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	9.00E-09
	A3WXE0_9BRAD	Putative uncharacterized protein (gene: NB311A_13581)	Nitrobacter sp. Nb-311A	1.70E-08
>GME_11572 - 6805: 8967 MW: 78051.65	F7SP90_9GAMM	Putative uncharacterized protein (gene: GME_11572)	Halomonas sp. TD01	0.00E+00
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.20E-28
	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	2.70E-24
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	9.60E-21
	C9WWY1_CLOBO	ORFX3 (Fragment)	Clostridium botulinum	3.80E-20
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	3.40E-18
	B2VCQ8_ERWT9	Putative uncharacterized protein (gene: ETA_30270)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	5.30E-18
	A8PMG2_9COXI	Putative uncharacterized protein (gene: RICGR_0717)	Rickettsiella grylli	2.00E-17
	C9WWY2_CLOBO	ORFX2	Clostridium botulinum	2.20E-12

>GME_11577 - 8991: 10388 MW: 49702.94	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	1.30E-306
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	4.40E-70
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	1.20E-68
	C5UY10_CLOBO	Toxin complex component ORF-X3 (gene: CLO_2647)	Clostridium botulinum E1 str. 'BoNT E Beluga'	4.10E-65
	C3KS19_CLOB6	Toxin complex component ORF-X3 (gene: CLJ_0010)	Clostridium botulinum (strain 657 / Type Ba4)	5.90E-62
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	4.60E-36
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	5.80E-34
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	1.70E-12
	Q45890_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	3.40E-09
	C3KS18_CLOB6	Toxin complex component ORF-X2 (gene: CLJ_0009)	Clostridium botulinum (strain 657 / Type Ba4)	2.10E-07
	D2T7Z2_ERWP6	Putative uncharacterized protein (gene: EPYR_03487)	Erwinia pyrifoliae (strain DSM 12163 / CIP 106111 / Ep16/96)	3.10E-07
>GME_11582 - 10418: 10933 MW: 19189.484	F7SP92_9GAMM	Putative uncharacterized protein (gene: GME_11582)	Halomonas sp. TD01	8.80E-113
	F7SP93_9GAMM	Putative uncharacterized protein (gene: GME_11587)	Halomonas sp. TD01	3.20E-50
	D2AVY8_STRRD	Putative uncharacterized protein (gene: Sros_0049)	Streptosporangium roseum (strain ATCC 12428 / DSM 43021 / JCM 3005 / NI 9100)	1.20E-48
	D8NKW8_RALSL	AidA (gene: aidA)	Ralstonia solanacearum CFBP2957	8.00E-47
	G2ZPF3_9RALS	Conserved hypothetical protein (gene: BDB_110385)	blood disease bacterium R229	3.80E-45

	B4EQ59_BURCJ	Nematocidal protein AidA (gene: aidA)	Burkholderia cepacia (strain J2315 / LMG 16656)	3.50E-35
	Q1I3W7_PSEE4	Putative uncharacterized protein (gene: PSEEN5035)	Pseudomonas entomophila (strain L48)	4.40E-33
	D9X2B6_STRVR	Nematocidal protein AidA (gene: SSQG_04103)	Streptomyces viridochromogenes DSM 40736	3.60E-31
	D3VHG3_XENNA	Putative uncharacterized protein (gene: XNC1_2550)	Xenorhabdus nematophila (strain ATCC 19061 / DSM 3370 / LMG 1036 / NCIB 9965 / AN6)	2.60E-19
>GME_11587 - 10993: 11508 MW: 18868.629	F7SP93_9GAMM	Putative uncharacterized protein (gene: GME_11587)	Halomonas sp. TD01	1.60E-111
	F7SP92_9GAMM	Putative uncharacterized protein (gene: GME_11582)	Halomonas sp. TD01	3.10E-50
	D2AVY8_STRRD	Putative uncharacterized protein (gene: Sros_0049)	Streptosporangium roseum (strain ATCC 12428 / DSM 43021 / JCM 3005 / NI 9100)	1.70E-34
	B4EQ59_BURCJ	Nematocidal protein AidA (gene: aidA)	Burkholderia cepacia (strain J2315 / LMG 16656)	3.00E-33
	D8NKW8_RALSL	AidA (gene: aidA)	Ralstonia solanacearum CFBP2957	2.60E-32
	D9X2B6_STRVR	Nematocidal protein AidA (gene: SSQG_04103)	Streptomyces viridochromogenes DSM 40736	8.90E-27
	Q4KD92_PSEF5	Putative uncharacterized protein (gene: PFL_2684)	Pseudomonas fluorescens (strain Pf-5 / ATCC BAA-477)	6.00E-19
	D3VHG3_XENNA	Putative uncharacterized protein (gene: XNC1_2550)	Xenorhabdus nematophila (strain ATCC 19061 / DSM 3370 / LMG 1036 / NCIB 9965 / AN6)	6.20E-12

	B6VM50_PHOAA	Similar to aida protein of ralstonia solanacearum (gene: aidA)	Photorhabdus asymbiotica subsp. asymbiotica (strain ATCC 43949 / 3105-77)	3.20E-10
>GME_11592 - 11533: 11889 MW: 13480.565	F7SP94_9GAMM	Putative uncharacterized protein (gene: GME_11592)	Halomonas sp. TD01	1.60E-73

Paenibacillus dendritiformis

Query	Match uniprot	Match protein name	Match protein species	Match score
>PDENDC454_25596 - 342: 761 MW: 16418.139	H3SNG9_9BACL	Toxin complex component ORF-X1 (gene: PDENDC454_25596)	Paenibacillus dendritiformis C454	1.20E-87
	Q6RI07_CLOBO	ORF-X1 (gene: orfx1)	Clostridium botulinum	4.30E-13
>PDENDC454_25591 - 784: 3051 MW: 85771.44	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	0.00E+00
	B0FNQ5_CLOBO	OrfX2	Clostridium botulinum	9.20E-139
	A8PMG2_9COXI	Putative uncharacterized protein (gene: RICGR_0717)	Rickettsiella grylli	3.10E-27
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	1.00E-26
	B2VCQ8_ERWT9	Putative uncharacterized protein (gene: ETA_30270)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	2.00E-26
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	1.60E-19
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	3.70E-17
	C4IHL5_CLOBU	Toxin complex component ORF-X3 (gene: CLP_2746)	Clostridium butyricum E4 str. BoNT E BL5262	3.90E-13
	C9WWY1_CLOBO	ORFX3 (Fragment)	Clostridium botulinum	6.30E-13

	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	7.10E-11
>COG0054 Riboflavin synthase beta-chain - 3162: 4643 MW: 55331.68	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	0.00E+00
	E9N6Q4_CLOBO	OrfX3 (gene: orfX3)	Clostridium botulinum	3.30E-135
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	3.10E-50
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	3.40E-35
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	1.40E-20
	C4IHL6_CLOBU	Toxin complex component ORF-X2 (gene: CLP_2747)	Clostridium butyricum E4 str. BoNT E BL5262	1.60E-16
	A3WXD6_9BRAD	Putative uncharacterized protein (gene: NB311A_13596)	Nitrobacter sp. Nb-311A	2.60E-15
	F7SP90_9GAMM	Putative uncharacterized protein (gene: GME_11572)	Halomonas sp. TD01	9.80E-15
	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	8.20E-11
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	2.20E-10
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	3.20E-08
	P71115_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	8.80E-06
>PDENDC454_25581 - 4667: 6072 MW: 52571.297	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	0.00E+00
	P71113_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	3.00E-81
	Q3SU90_NITWN	Putative uncharacterized protein (gene: Nwi_0887)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	9.00E-38

	D2TXI9_9ENTR	Putative uncharacterized protein (gene: ARN_081010)	Arsenophonus nasoniae	3.60E-36
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	7.50E-32
	B2VCQ7_ERWT9	Putative uncharacterized protein (gene: ETA_30260)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	5.70E-18
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	2.30E-08
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	3.40E-08

Nitrobacter winogradskyi

Query	Match uniprot	Match protein name	Match protein species	Match score
>Nwi_0887 - 2435137: 2436387 MW: 44933.914	Q3SU90_NITWN	Putative uncharacterized protein (gene: Nwi_0887)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	4.40E-278
	A3WXE0_9BRAD	Putative uncharacterized protein (gene: NB311A_13581)	Nitrobacter sp. Nb-311A	3.00E-249
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	2.00E-38
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	8.40E-38
	C3KS15_CLOB6	P-47 protein (gene: p47)	Clostridium botulinum (strain 657 / Type Ba4)	4.10E-25
	B0FNR1_CLOBO	OrfX3 (Fragment)	Clostridium botulinum	6.50E-12
	B2VCQ7_ERWT9	Putative uncharacterized protein (gene: ETA_30260)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	1.50E-11
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	4.60E-11
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	7.00E-11

>Nwi_0886 - 2436442: 2438055 MW: 56572.258	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	0.00E+00
	A3WXE1_9BRAD	Putative uncharacterized protein (gene: NB311A_13586)	Nitrobacter sp. Nb-311A	0.00E+00
	Q3SU92_NITWN	Putative uncharacterized protein (gene: Nwi_0885)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	7.90E-107
	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	7.10E-69
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	1.10E-68
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	1.60E-50
	E9N6Q4_CLOBO	OrfX3 (gene: orfX3)	Clostridium botulinum	3.50E-46
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	1.60E-27
	Q6RI02_CLOBO	ORF-X2 (gene: orfx2)	Clostridium botulinum	3.10E-21
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	4.30E-21
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	2.00E-20
	P71115_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	7.90E-13
>Nwi_0885 - 2438058: 2439650 MW: 56652.863	Q3SU92_NITWN	Putative uncharacterized protein (gene: Nwi_0885)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	0.00E+00
	A3WXD5_9BRAD	Putative uncharacterized protein (gene: NB311A_13591)	Nitrobacter sp. Nb-311A	0.00E+00
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	1.70E-106
	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	4.30E-59
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	6.70E-57

	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	1.10E-43
	Q6RI09_CLOBO	ORF-X3 (gene: orfx3)	Clostridium botulinum	6.80E-33
	F7SP90_9GAMM	Putative uncharacterized protein (gene: GME_11572)	Halomonas sp. TD01	1.30E-24
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	3.00E-23
	B2VCQ8_ERWT9	Putative uncharacterized protein (gene: ETA_30270)	Erwinia tasmaniensis (strain DSM 17950 / Et1/99)	3.10E-16
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	2.90E-15
	P71115_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	3.80E-07
>Nwi_0884 - 2439745: 2441265 MW: 52274.707	Q3SU93_NITWN	Putative uncharacterized protein (gene: Nwi_0884)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	0.00E+00
	A3WXD6_9BRAD	Putative uncharacterized protein (gene: NB311A_13596)	Nitrobacter sp. Nb-311A	5.30E-300
	A3WXD5_9BRAD	Putative uncharacterized protein (gene: NB311A_13591)	Nitrobacter sp. Nb-311A	1.80E-44
	Q3SU92_NITWN	Putative uncharacterized protein (gene: Nwi_0885)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.30E-43
	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	1.80E-23
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	1.30E-17
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	4.80E-12
	C4IHL5_CLOBU	Toxin complex component ORF-X3 (gene: CLP_2746)	Clostridium butyricum E4 str. BoNT E BL5262	9.20E-07
	C5UY10_CLOBO	Toxin complex component ORF-X3 (gene: CLO_2647)	Clostridium botulinum E1 str. 'BoNT E Beluga'	9.20E-07
	B0FNR1_CLOBO	OrfX3 (Fragment)	Clostridium botulinum	3.30E-06

	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	3.50E-06
	E9N6Q4_CLOBO	OrfX3 (gene: orfX3)	Clostridium botulinum	5.60E-06
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	1.10E-05
	F7SP90_9GAMM	Putative uncharacterized protein (gene: GME_11572)	Halomonas sp. TD01	1.70E-05

Streptomyces scabei

Query	Match uniprot	Match protein name	Match protein species	Match score
>with SCAB58511 - 3636359: 3638053 MW: 59991.6	Q3SU92_NITWN	Putative uncharacterized protein (gene: Nwi_0885)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.00E-47
	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	2.70E-39
	Q6RI03_CLOBO	ORF-X3 (gene: orfx3)	Clostridium botulinum	3.90E-30
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	1.40E-28
	F7SP90_9GAMM	Putative uncharacterized protein (gene: GME_11572)	Halomonas sp. TD01	2.20E-17
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	4.60E-15
	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	2.10E-12
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	8.70E-07
	D2TXI8_9ENTR	Toxin complex component ORF-X2 (gene: ARN_08100)	Arsenophonus nasoniae	7.20E-05
>with SCAB58531 - 3638202: 3639821 MW: 57924.875	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	5.00E-34
	Q3SU92_NITWN	Putative uncharacterized protein (gene: Nwi_0885)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.10E-30
	B0FNR1_CLOBO	OrfX3 (Fragment)	Clostridium botulinum	2.10E-21

	C3KS18_CLOB6	Toxin complex component ORF-X2 (gene: CLJ_0009)	Clostridium botulinum (strain 657 / Type Ba4)	1.10E-09
	C3KS15_CLOB6	P-47 protein (gene: p47)	Clostridium botulinum (strain 657 / Type Ba4)	2.80E-05

***Achromobacter xylosoxidans* AXX-A**

Query	Match uniprot	Match protein name	Match protein species	Match score
>AXXA_06258 - 163885: 165699 MW: 64139.195	F7SWZ6_ALCXX	Putative uncharacterized protein (gene: AXXA_06258)	Achromobacter xylosoxidans AXX-A	0.00E+00
	A3WXE1_9BRAD	Putative uncharacterized protein (gene: NB311A_13586)	Nitrobacter sp. Nb-311A	1.20E-25
	E9N6Q4_CLOBO	OrfX3 (gene: orfX3)	Clostridium botulinum	1.80E-15
	H3SNG8_9BACL	Toxin complex component ORF-X2 (gene: PDENDC454_25591)	Paenibacillus dendritiformis C454	1.30E-14
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	7.10E-14
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	5.40E-13
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	4.20E-07
	P71115_CLOBO	P-47 protein (gene: P-47)	Clostridium botulinum	5.00E-04
>AXXA_06253 - 161108: 163804 MW: 96362.31	F7SWZ5_ALCXX	Putative uncharacterized protein (gene: AXXA_06253)	Achromobacter xylosoxidans AXX-A	0.00E+00
	F7SWZ6_ALCXX	Putative uncharacterized protein (gene: AXXA_06258)	Achromobacter xylosoxidans AXX-A	9.10E-04
>AXXA_06263 - 165738: 167090 MW: 48627.824	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	2.90E-301
	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	9.90E-37
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	9.70E-27

	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	1.60E-25
	A7GBF7_CLOBL	Toxin complex component ORF-X3 (gene: CLI_0845)	Clostridium botulinum (strain Langeland / NCTC 10281 / Type F)	2.00E-24
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	3.30E-13
	Q45843_CLOBO	P-47 protein (Fragment) (gene: p-47)	Clostridium botulinum	2.10E-06
	E9N6Q5_CLOBO	OrfX2 (gene: orfX2)	Clostridium botulinum	7.70E-04
>AXXA_06268 - 167087: 168436 MW: 47213.496	F7SWZ8_ALCXX	Putative uncharacterized protein (gene: AXXA_06268)	Achromobacter xylosoxidans AXX-A	2.30E-297
	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	1.70E-04
	A3WXD5_9BRAD	Putative uncharacterized protein (gene: NB311A_13591)	Nitrobacter sp. Nb-311A	2.90E-04
	Q3MB33_ANAVT	Peptidase C14, caspase catalytic subunit p20 (gene: Ava_2183)	Anabaena variabilis (strain ATCC 29413 / PCC 7937)	1.00E-03
	Q3SU92_NITWN	Putative uncharacterized protein (gene: Nwi_0885)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	7.40E-03

Pseudomonas putida				
Query	Match uniprot	Match protein name	Match protein species	Match score
>similar to GB:D14445, SP:Q07411, and PID:286035; i - 3903288: 3904520 MW: 44707.453	Q88LC8_PSEPK	P-47-related protein (gene: PP_2007)	Pseudomonas putida (strain KT2440)	2.70E-282
	Q3SU90_NITWN	Putative uncharacterized protein (gene: Nwi_0887)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	2.10E-39
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	7.80E-27
	Q6RI05_CLOBO	P-47 (gene: p47)	Clostridium botulinum	1.10E-21

	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	5.70E-20
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	1.60E-13
	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	1.60E-13
	A8PMG3_9COXI	Putative uncharacterized protein (gene: RICGR_0718)	Rickettsiella grylli	1.90E-11
	D2TXI9_9ENTR	Putative uncharacterized protein (gene: ARN_081010)	Arsenophonus nasoniae	1.60E-07
>PP_2006 - 3904411: 3906189 MW: 63712.24	Q88LC9_PSEPK	Putative uncharacterized protein (gene: PP_2006)	Pseudomonas putida (strain KT2440)	0.00E+00
	Q3SU91_NITWN	Putative uncharacterized protein (gene: Nwi_0886)	Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391)	4.90E-68
	E9N6Q4_CLOBO	OrfX3 (gene: orfX3)	Clostridium botulinum	8.40E-36
	H3SNG7_9BACL	Toxin complex component ORF-X3 (gene: PDENDC454_25586)	Paenibacillus dendritiformis C454	1.80E-35
	F7SP91_9GAMM	Putative uncharacterized protein (gene: GME_11577)	Halomonas sp. TD01	5.60E-34
	F7SWZ7_ALCXX	Toxin complex component ORF-X3 (gene: AXXA_06263)	Achromobacter xylosoxidans AXX-A	3.60E-25
	E5B9Q4_ERWAM	Putative uncharacterized protein (gene: EAIL5_3311)	Erwinia amylovora ATCC BAA-2158	5.80E-19
	Q6RI02_CLOBO	ORF-X2 (gene: orfx2)	Clostridium botulinum	3.80E-17
	H3SNG6_9BACL	P-47 protein (Fragment) (gene: PDENDC454_25581)	Paenibacillus dendritiformis C454	5.50E-15
>similar to GP:9949202; identified by sequence simi - 3906378: 3907496 MW: 41070.066	E4RBH0_PSEPB	Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase (gene: PPUBIRD1_3665)	Pseudomonas putida (strain BIRD-1)	4.90E-250

>identified by match to PFAM protein family HMM PF0 - 3907501: 3908541 MW: 38701.61	Q88LD1_PSEPK	Transcriptional regulator, AraC family (gene: PP_2004)	Pseudomonas putida (strain KT2440)	4.90E-228

Appendix Table 2: Proteins identified in the culture supernatant of *C. botulinum* A1 19397 at 24 h only

Row number	Uniprot accession	Locus	GI number	Common name	Main functional role	Functional sub role
1	A7FPK3	CLB_3627	153932124	stage V sporulation protein G	Cellular processes	Sporulation and germination
2	A7FPR1	CLB_2579	153931144	putative manganese-dependent inorganic pyrophosphatase	Central intermediary metabolism	Phosphorus compounds
3	A7FQ40	CLB_3520	153931344	30S ribosomal protein S5	Protein synthesis	Ribosomal proteins: synthesis and modification
4	A7FQ93	CLB_0100	153931071	conserved hypothetical protein	Hypothetical proteins	Conserved
5	A7FQ94	CLB_0101	153931499	arginine deiminase	Energy metabolism	Amino acids and amines
6	A7FQD5	CLB_0142	153930826	radical SAM domain protein	Unknown function	Enzymes of unknown specificity
7	A7FQI8	CLB_0208	153932206	S-adenosylmethionine synthetase	Central intermediary metabolism	Other
8	A7FQJ2	CLB_0212	153933462	ribosomal subunit interface protein	Protein synthesis	Translation factors
9	A7FQM3	CLB_0254	153931001	glutamyl aminopeptidase family protein	Protein fate	Degradation of proteins, peptides, and glycopeptides
10	A7FR63	CLB_0461	153933913	ABC transporter, ATP-binding protein	Transport and binding proteins	Unknown substrate
11	A7FRB8	CLB_0545	153932683	NlpC/P60 family protein	Unknown function	General
12	A7FS27	CLB_0812	153934103	putative thiamine pyridinylase I	Central intermediary metabolism	Other
13	A7FSA3	CLB_0893	153933318	putative NADPH-dependent FMN reductase	Energy metabolism	Electron transport
14	A7FSI1	CLB_0974	153933208	nitroreductase family protein	Unknown function	Enzymes of unknown specificity
15	A7FSL0	CLB_1003	153933882	putative anion ABC transporter, solute-binding protein	Transport and binding proteins	Anions
16	A7FSR1	CLB_1054	153931801	pyridoxal-phosphate-dependent aminotransferase family	Amino acid biosynthesis	Aspartate family
17	A7FT23	CLB_1170	153933706	conserved hypothetical protein	Hypothetical proteins	Conserved
18	A7FTD9	CLB_1288	153931831	thioredoxin family protein	Energy metabolism	Electron transport

19	A7FTE2	CLB_1292	153933934	glycine reductase complex component C, alpha subunit	Energy metabolism	Amino acids and amines
20	A7FTG3	CLB_1313	153931205	xanthine dehydrogenase family protein, molybdopterin-binding subunit	Unknown function	Enzymes of unknown specificity
21	A7FTG4	CLB_1314	153932992	xanthine dehydrogenase family protein, FAD-binding subunit	Unknown function	Enzymes of unknown specificity
22	A7FTH2	CLB_1322	153933923	RNA-metabolising metallo-beta-lactamase	Transcription	Other
23	A7FTQ9	CLB_1409	153932613	nitroreductase family protein	Unknown function	Enzymes of unknown specificity
24	A7FTR1	CLB_1411	153931072	major cold shock protein CspA	Cellular processes	Adaptations to atypical conditions
25	A7FU00	CLB_1507	153931523	molybdate ABC transporter, periplasmic molybdate-binding protein	Transport and binding proteins	Anions
26	A7FU03	CLB_1510	153932021	conserved hypothetical protein	Hypothetical proteins	Conserved
27	A7FU16	CLB_1523	153933046	lipoprotein, NLPA family	Cell envelope	Other
28	A7FU90	CLB_1600	153933339	pyridine nucleotide-disulphide oxidoreductase family protein	Unknown function	Enzymes of unknown specificity
29	A7FUB1	CLB_1622	153933474	putative glucokinase	Energy metabolism	Glycolysis/gluconeogenesis
30	A7FUI2	CLB_1693	153931747	conserved hypothetical protein	Hypothetical proteins	Conserved
31	A7FUL4	CLB_1739	153932470	aspartate aminotransferase	Amino acid biosynthesis	Aspartate family
32	A7FUM4	CLB_1749	153932416	conserved hypothetical protein	Hypothetical proteins	Conserved
33	A7FUS3	CLB_1804	153930854	hydrolase, NUDIX family	Unknown function	Enzymes of unknown specificity
34	A7FUX2	CLB_1854	153931842	glutamate decarboxylase	Energy metabolism	Amino acids and amines
35	A7FV42	CLB_1925	153932933	heat shock protein	Protein fate	Protein folding and stabilization
36	A7FVB3	CLB_2000	153932535	triacylglycerol lipase	Fatty acid and phospholipid metabolism	Degradation
37	A7FW59	CLB_2366	153931465	pantetheine-phosphate adenylyltransferase	Biosynthesis of cofactors, prosthetic groups, and carriers	Pantothenate and coenzyme A

38	A7FWD9	CLB_2446	153931359	putative phage capsid protein	Mobile and extrachromosomal element functions	Prophage functions
39	A7FWK5	CLB_2515	153932795	selenium metabolism protein YedF	Unknown function	General
40	A7FWK6	CLB_2516	153932049	putative alanine dehydrogenase	Energy metabolism	Amino acids and amines
41	A7FWM6	CLB_2536	153932337	serine hydroxymethyltransferase	Amino acid biosynthesis	Serine family
42	A7FWZ8	CLB_2689	153933380	chemotaxis protein CheA	Cellular processes	Chemotaxis and motility
43	A7FX19	CLB_2710	153932262	DJ-1 family protein	Unknown function	General
44	A7FXC4	CLB_2829	153931216	6,7-dimethyl-8-ribityllumazine synthase	Biosynthesis of cofactors, prosthetic groups, and carriers	Riboflavin, FMN, and FAD
45	A7FXC8	CLB_2833	153934160	aminoacyl-histidine dipeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
46	A7FXE2	CLB_2847	153933697	amidohydrolase family protein	Unknown function	Enzymes of unknown specificity
47	A7FXG8	CLB_2873	153932266	conserved hypothetical protein TIGR00106	Hypothetical proteins	Conserved
48	A7FXG9	CLB_2874	153934268	thiamine biosynthesis protein ThiC	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine
49	A7FXL0	CLB_2918	153932252	HIT family protein	Unknown function	General
50	A7FXY6	CLB_3061	153931153	conserved hypothetical protein	Hypothetical proteins	Conserved
51	A7FY60	CLB_3140	153931954	putative N-acetylmuramoyl-L-alanine amidase	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides
52	A7FYJ2	CLB_3278	153933721	aspartate carbamoyltransferase	Purines, pyrimidines, nucleosides, and nucleotides	Pyrimidine ribonucleotide biosynthesis
53	A7FYN5	CLB_3346	153931140	R-phenyllactate dehydratase, B subunit	Energy metabolism	Amino acids and amines
54	A7FYT4	CLB_3396	153931339	zinc metalloprotease, aminopeptidase I family	Protein fate	Degradation of proteins, peptides, and glycopeptides
55	A7FZ16	CLB_3480	153934022	putative 2-oxoacid:acceptor oxidoreductase, delta subunit	Energy metabolism	Other

56	A7FZ57	CLB_3525	153931371	50S ribosomal protein L5	Protein synthesis	Ribosomal proteins: synthesis and modification
57	A7FZ64	CLB_3532	153932380	50S ribosomal protein L22	Protein synthesis	Ribosomal proteins: synthesis and modification
58	A7FZ67	CLB_3535	153933049	50S ribosomal protein L23	Protein synthesis	Ribosomal proteins: synthesis and modification
59	A7FZ80	CLB_3548	153932159	50S ribosomal protein L1	Protein synthesis	Ribosomal proteins: synthesis and modification
60	A7FZ92	CLB_3562	153930911	threonine dehydratase	Amino acid biosynthesis	Pyruvate family
61	A7FZC1	CLB_3616	153932845	MazG family protein	Unknown function	General
62	A7FZC7	CLB_3647	153933851	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan
63	A7FZH2	CLB_3722	153933289	2-hydroxyglutaryl-CoA dehydratase, D-component	Energy metabolism	Amino acids and amines

	Replicate A					Replicate B		
Row number	Protein identification probability	Exclusive unique peptide count	Exclusive unique spectrum count	Total spectrum count	Protein coverage (%)	Protein identification probability	Exclusive unique peptide count	Row number
1	99.6%	2	2	7	24.2%	99.3%	2	1
2	100.0%	6	6	17	13.0%	96.9%	2	2
3	99.0%	2	2	6	10.9%	99.2%	2	3
4	99.0%	2	2	2	12.3%	97.8%	2	4
5	100.0%	8	8	30	23.8%	28.8%	1	5
6	100.0%	4	4	15	9.6%	100.0%	4	6
7	99.8%	2	2	6	7.2%	50.0%	1	7
8	50.0%	1	1	19	5.1%	50.0%	1	8
9	50.0%	1	1	1	3.5%	98.5%	2	9
10	9.1%	1	1	1	4.2%	95.3%	2	10
11	100.0%	4	4	10	7.0%	100.0%	4	11
12	100.0%	8	9	28	22.8%	100.0%	6	12
13	25.5%	1	1	1	6.0%	43.6%	1	13
14	8.1%	1	1	1	8.1%	98.8%	2	14
15	97.8%	2	2	3	11.3%	100.0%	5	15
16	99.7%	2	2	3	7.1%	100.0%	3	16
17	32.2%	1	1	1	6.6%	50.0%	1	17
18	99.0%	2	2	6	15.1%	98.2%	2	18

19	100.0%	4	4	9	13.7%	98.8%	3	19
20	96.1%	2	2	2	3.4%	97.8%	2	20
21	41.2%	1	1	1	3.4%	99.8%	2	21
22	100.0%	4	4	9	8.3%	31.0%	1	22
23	32.5%	1	2	4	6.0%	50.0%	1	23
24	100.0%	3	3	20	46.3%	100.0%	3	24
25	100.0%	4	4	10	20.3%	81.5%	2	25
26	100.0%	4	4	5	29.6%	100.0%	3	26
27	100.0%	3	3	13	13.6%	99.8%	2	27
28	50.0%	1	1	4	1.6%	99.9%	3	28
29	36.1%	1	1	1	4.1%	50.0%	1	29
30	100.0%	4	4	12	16.8%	100.0%	4	30
31	99.8%	2	2	4	5.6%	99.8%	2	31
32	100.0%	10	10	61	25.9%	19.6%	1	32
33	7.1%	1	1	1	6.2%	13.5%	1	33
34	50.0%	1	1	3	2.6%	50.0%	1	34
35	100.0%	5	5	19	9.3%	100.0%	5	35
36	99.8%	2	2	11	6.3%	50.0%	1	36
37	100.0%	3	3	4	20.7%	99.8%	3	37

38	99.8%	2	2	8	6.9%	50.0%	1	38
39	21.7%	1	1	2	6.7%	99.3%	2	39
40	13.7%	1	1	1	3.3%	10.4%	1	40
41	98.8%	2	2	4	5.3%	23.6%	1	41
42	9.5%	1	1	1	1.4%	99.5%	2	42
43	100.0%	3	3	6	20.8%	100.0%	3	43
44	50.0%	1	1	2	6.5%	100.0%	3	44
45	99.8%	2	2	5	5.0%	99.8%	2	45
46	100.0%	4	4	10	14.9%	100.0%	4	46
47	98.5%	2	2	3	25.0%	97.4%	2	47
48	100.0%	6	7	14	14.0%	100.0%	7	48
49	99.9%	3	4	8	30.7%	50.0%	1	49
50	97.9%	2	2	4	7.8%	100.0%	3	50
51	99.3%	2	2	4	2.5%	50.0%	1	51
52	100.0%	7	7	22	26.7%	99.8%	2	52
53	97.4%	2	2	2	5.7%	87.6%	2	53
54	99.8%	2	2	7	6.3%	50.0%	1	54
55	50.0%	1	1	3	15.9%	99.3%	2	55

56	99.1%	2	2	3	10.6%	97.6%	2	56
57	42.8%	1	1	4	9.0%	99.8%	2	57
58	98.6%	2	2	3	18.6%	49.2%	1	58
59	100.0%	4	5	11	18.8%	100.0%	3	59
60	99.9%	3	4	9	10.1%	98.2%	2	60
61	50.0%	1	1	1	2.3%	99.8%	2	61
62	100.0%	6	7	19	17.3%	98.9%	2	62
63	96.2%	2	2	5	760.0%	100.0%	4	63

Replicate C

Exclusive unique spectrum count	Total spectrum count	Protein coverage (%)	Protein identification probability	Exclusive unique peptide count	Exclusive unique spectrum count	Total spectrum count	Protein coverage (%)
3	11	24.2%	50.0%	1	2	5	15.8%
2	3	3.5%	99.8%	2	2	3	5.1%
2	6	10.9%	7.0%	1	1	1	5.5%
2	3	12.3%	97.4%	2	2	3	12.3%
1	2	2.2%	100.0%	7	7	15	21.3%
4	11	9.6%	100.0%	4	4	14	9.6%
1	3	3.3%	99.7%	2	2	4	7.2%
1	35	5.1%	99.0%	2	3	58	15.4%
2	2	6.1%	100.0%	3	3	4	9.9%
2	2	8.4%	100.0%	3	3	4	13.0%
4	17	7.0%	99.8%	2	2	4	2.8%
6	9	17.3%	100.0%	5	6	10	15.1%
1	1	6.0%	98.6%	2	2	3	12.0%
2	3	13.3%	97.2%	2	2	3	13.3%
5	10	21.1%	100.0%	7	8	17	31.2%
3	5	10.2%	100.0%	4	4	9	12.2%
1	3	6.6%	99.8%	2	2	4	12.0%
2	4	15.1%	20.2%	1	1	2	7.5%

3	3	9.5%	100.0%	4	4	7	12.6%
2	4	3.1%	99.8%	2	2	3	3.4%
2	3	7.5%	99.8%	2	2	4	7.5%
1	1	1.8%	16.2%	1	1	1	1.8%
2	11	6.0%	99.4%	2	3	7	11.5%
3	23	46.3%	50.0%	1	1	2	16.4%
2	2	7.7%	99.6%	3	3	5	13.7%
3	8	20.4%	99.8%	2	2	4	16.1%
2	4	8.5%	100.0%	3	3	4	13.6%
3	5	4.8%	99.7%	3	3	5	5.7%
1	3	4.1%	99.5%	2	2	5	6.7%
4	11	17.1%	6.6%	1	1	1	2.8%
2	3	5.6%	99.9%	3	4	9	7.8%
1	1	2.3%	100.0%	9	10	30	23.1%
1	1	7.3%	100.0%	4	4	7	24.2%
1	3	2.6%	99.8%	2	2	4	5.4%
5	12	9.3%	93.4%	2	2	3	4.2%
1	2	3.3%	50.0%	1	1	2	3.3%
3	7	20.7%	93.6%	2	2	3	12.8%

1	2	4.0%	99.6%	2	2	4	6.9%
3	5	13.9%	50.0%	1	1	2	6.7%
1	2	3.0%	99.8%	2	2	4	7.3%
1	2	2.7%	99.8%	2	2	4	5.3%
2	6	2.9%	99.7%	2	2	6	2.9%
3	7	20.8%	100.0%	3	3	3	20.8%
3	6	26.6%	100.0%	3	3	4	22.1%
2	3	5.0%	100.0%	4	4	11	9.4%
4	6	14.9%	99.8%	2	2	3	7.2%
2	4	25.0%	98.8%	2	2	5	25.0%
7	11	16.0%	100.0%	5	5	6	11.7%
1	3	8.8%	99.2%	2	2	4	21.9%
3	6	11.5%	99.4%	2	2	4	7.8%
1	4	1.4%	99.3%	2	2	4	2.5%
2	4	7.8%	100.0%	7	7	23	26.7%
2	2	5.7%	99.8%	2	2	5	5.7%
1	3	3.0%	100.0%	4	4	16	11.8%
2	12	31.9%	98.9%	2	2	8	31.9%

2	4	10.6%	50.0%	1	1	1	6.1%
2	5	17.1%	23.2%	1	1	3	9.0%
1	3	10.3%	39.7%	1	1	3	10.3%
3	13	14.4%	99.8%	2	2	3	10.0%
2	4	6.9%	99.8%	3	4	7	10.1%
2	4	4.8%	50.0%	1	1	1	2.5%
2	3	4.8%	100.0%	3	3	5	9.4%
5	7	1340.0%	100.0%	5	5	12	1520.0%

Appendix table 3: Proteins identified in the supernatant of *C. botulinum* A1 19397 at 24 h and 96 h

Row number	Uniprot accession	Locus	GI number	Common name	Main functional role	Functional sub role
1	A7FPG2	CLB_0480	153931527	peptidase T	Protein fate	Degradation of proteins, peptides, and glycopeptides
2	A7FPK6	CLB_0049	153933357	conserved hypothetical protein	Hypothetical proteins	Conserved
3	A7FPN7	CLB_3677	153932539	beta-hydroxyacyl-(acyl-carrier-protein) dehydratase FabZ	Fatty acid and phospholipid metabolism	Biosynthesis
4	A7FPQ2	CLB_0234	153933873	fructose-1,6-bisphosphate aldolase, class II	Energy metabolism	Glycolysis/gluconeogenesis
5	A7FPR3	CLB_2581	153933262	flagellar basal body rod protein	Cellular processes	Chemotaxis and motility
6	A7FPY3	CLB_2134	153932288	acyl-CoA dehydrogenase family protein	Fatty acid and phospholipid metabolism	Degradation
7	A7FPY4	CLB_2135	153931286	putative (R)-2-hydroxyglutaryl-CoA dehydratase, beta subunit	Energy metabolism	Amino acids and amines
8	A7FPY5	CLB_2136	153932215	putative (R)-2-hydroxyglutaryl-CoA dehydratase, alpha subunit	Energy metabolism	Amino acids and amines
9	A7FPZ1	CLB_2716	153932999	conserved hypothetical protein	Hypothetical proteins	Conserved
10	A7FPZ5	CLB_2296	153933414	ribosome recycling factor	Protein synthesis	Translation factors
11	A7FPZ6	CLB_2297	153932400	UMP kinase	Purines, pyrimidines, nucleosides, and nucleotides	Nucleotide and nucleoside interconversions
12	A7FPZ7	CLB_2298	153931043	translation elongation factor Ts	Protein synthesis	Translation factors
13	A7FQ63	CLB_0039	153931933	methionine gamma-lyase	Energy metabolism	Amino acids and amines
14	A7FQ86	CLB_0093	153930936	PSP1 domain protein	Unknown function	General
15	A7FQ97	CLB_0104	153932306	polysaccharide deacetylase family protein	Energy metabolism	Biosynthesis and degradation of polysaccharides
16	A7FQA4	CLB_0111	153931420	methionyl-tRNA synthetase	Protein synthesis	tRNA aminoacylation
17	A7FQC8	CLB_0135	153932381	isoleucyl-tRNA synthetase	Protein synthesis	tRNA aminoacylation

18	A7FQN6	CLB_0267	153930852	glyceraldehyde-3-phosphate dehydrogenase, type I	Energy metabolism	Glycolysis/gluconeogenesis
19	A7FQN7	CLB_0268	153931900	phosphoglycerate kinase	Energy metabolism	Glycolysis/gluconeogenesis
20	A7FQN8	CLB_0269	153932906	triosephosphate isomerase	Energy metabolism	Glycolysis/gluconeogenesis
21	A7FQP0	CLB_0271	153932900	enolase	Energy metabolism	Glycolysis/gluconeogenesis
22	A7FQZ7	CLB_0387	153931946	metallopeptidase, family M24	Protein fate	Degradation of proteins, peptides, and glycopeptides
23	A7FQZ8	CLB_0388	153932445	aldehyde-alcohol dehydrogenase	Energy metabolism	Fermentation
24	A7FR44	CLB_0434	153931423	clpB protein	Protein fate	Protein folding and stabilization
25	A7FRL6	CLB_0645	153932108	putative lipoprotein	Cell envelope	Other
26	A7FS45	CLB_0830	153933679	conserved hypothetical protein	Hypothetical proteins	Conserved
27	A7FS58	CLB_0843	153932893	hemagglutinin component HA70	Cellular processes	Pathogenesis
28	A7FS59	CLB_0844	153933825	hemagglutinin component HA17	Cellular processes	Pathogenesis
29	A7FS60	CLB_0845	153932677	hemagglutinin component HA33	Cellular processes	Pathogenesis
30	A7FS62	CLB_0847	153932404	botulinum neurotoxin type A1, nontoxic-nonhemagglutinin component, NTNH	Cellular processes	Pathogenesis
31	A7FS63	CLB_0848	153931567	bontoxilysin A	Cellular processes	Pathogenesis
32	A7FSQ9	CLB_1052	153933200	transcriptional regulator, MarR family	Regulatory functions	DNA interactions
33	A7FSR0	CLB_1053	153934318	putative cyclase	Unknown function	General
34	A7FSX7	CLB_1122	153932796	putative glycosyl hydrolase	Energy metabolism	Biosynthesis and degradation of polysaccharides
35	A7FT31	CLB_1178	153931753	putative thiosulfate sulfurtransferase	Central intermediary metabolism	Sulfur metabolism
36	A7FTB6	CLB_1264	153933438	deoxyuridine 5'-triphosphate nucleotidohydrolase	Purines, pyrimidines, nucleosides, and nucleotides	2'-Deoxyribonucleotide metabolism
37	A7FTD6	CLB_1284	153933170	glycine reductase, subunits ABC	Energy metabolism	Amino acids and amines
38	A7FTD7	CLB_1285	153931631	glycine reductase complex component B, gamma subunit; selenocysteine-containing	Energy metabolism	Amino acids and amines
39	A7FTE1	CLB_1291	153933177	glycine reductase complex component C, beta subunit	Energy metabolism	Amino acids and amines

40	A7FTH1	CLB_1321	153932009	transaldolase	Energy metabolism	Pentose phosphate pathway
41	A7FTI3	CLB_1333	153930819	copper chaperone CopZ	Cellular processes	Detoxification
42	A7FTK2	CLB_1352	153931592	rubrerythrin	Energy metabolism	Electron transport
43	A7FTT2	CLB_1432	153934100	NADH-dependent butanol dehydrogenase	Energy metabolism	Fermentation
44	A7FTW3	CLB_1463	153933317	alcohol dehydrogenase, iron-containing	Energy metabolism	Fermentation
45	A7FTW9	CLB_1470	153932249	thermolysin metallopeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
46	A7FTX0	CLB_1471	153931929	thermolysin metallopeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
47	A7FU34	CLB_1542	153933395	redox family protein	Unknown function	General
48	A7FU67	CLB_1577	153932015	GTP cyclohydrolase I	Biosynthesis of cofactors, prosthetic groups, and carriers	Folic acid
49	A7FU97	CLB_1607	153932082	arginine deiminase	Energy metabolism	Amino acids and amines
50	A7FUC7	CLB_1638	153931218	collagenase	Protein fate	Degradation of proteins, peptides, and glycopeptides
51	A7FUE9	CLB_1660	153931169	NADH-dependent butanol dehydrogenase	Energy metabolism	Fermentation
52	A7FUF4	CLB_1665	153933878	Ig group 2 domain protein	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides
53	A7FUG1	CLB_1672	153933598	conserved hypothetical protein	Hypothetical proteins	Conserved
54	A7FUJ2	CLB_1717	153933744	leucine rich repeat protein	Unknown function	General
55	A7FUJ4	CLB_1719	153933365	ferritin family protein	Transport and binding proteins	Cations and iron carrying compounds
56	A7FUL2	CLB_1737	153934333	aminotransferase, classes I and II	Unknown function	Enzymes of unknown specificity
57	A7FUM1	CLB_1746	153932116	glutamate dehydrogenase, NAD-specific	Energy metabolism	Amino acids and amines
58	A7FUR9	CLB_1800	153930886	purine nucleoside phosphorylase I, inosine and guanosine-specific	Purines, pyrimidines, nucleosides, and nucleotides	Salvage of nucleosides and nucleotides
59	A7FUS5	CLB_1806	153932435	nlpC/P60 family protein	Unknown function	General
60	A7FUX5	CLB_1857	153932072	clostripain	Cellular processes	Pathogenesis

61	A7FV13	CLB_1896	153931951	leucine-rich repeat protein	Unknown function	General
62	A7FV14	CLB_1897	153931198	carbohydrate binding protein	Unknown function	General
63	A7FV99	CLB_1984	153933768	myosin-cross-reactive antigen family protein	Cell envelope	Surface structures
64	A7FVD3	CLB_2021	153931181	conserved hypothetical protein	Hypothetical proteins	Conserved
65	A7FVH5	CLB_2072	153932966	LPXTG-motif cell wall anchor domain	Cell envelope	Other
66	A7FVJ3	CLB_2090	153931782	putative PTS system, L-Ascorbate family, IIB component	Signal transduction	PTS
67	A7FVT1	CLB_2205	153931971	metallopeptidase, family M24	Protein fate	Degradation of proteins, peptides, and glycopeptides
68	A7FVX2	CLB_2261	153933969	phosphocarrier protein HPr	Signal transduction	PTS
69	A7FW29	CLB_2330	153933030	D-proline reductase, PrdA proprotein	Energy metabolism	Amino acids and amines
70	A7FW33	CLB_2337	153931124	conserved hypothetical protein	Hypothetical proteins	Conserved
71	A7FW39	CLB_2343	153932921	proline racemase	Energy metabolism	Amino acids and amines
72	A7FW44	CLB_2349	153931385	D-proline reductase, PrdA proprotein	Energy metabolism	Amino acids and amines
73	A7FW63	CLB_2370	153933450	conserved hypothetical protein	Hypothetical proteins	Conserved
74	A7FW76	CLB_2383	153931783	peptide deformylase	Protein fate	Protein modification and repair
75	A7FWM4	CLB_2534	153932030	ornithine carbamoyltransferase	Amino acid biosynthesis	Glutamate family
76	A7FWM5	CLB_2535	153931097	carbamate kinase	Energy metabolism	Amino acids and amines
77	A7FWN6	CLB_2546	153933857	peptidase, M20/M25/M40 family	Protein fate	Degradation of proteins, peptides, and glycopeptides
78	A7FWP7	CLB_2557	153932911	putative thiosulfate sulfurtransferase	Central intermediary metabolism	Sulfur metabolism
79	A7FWQ1	CLB_2561	153934176	metallo-beta-lactamase family protein/flavodoxin	Energy metabolism	Electron transport
80	A7FWQ4	CLB_2564	153933577	peptide deformylase	Protein fate	Protein modification and repair
81	A7FWQ6	CLB_2566	153932485	V-type sodium ATPase, B subunit	Energy metabolism	ATP-proton motive force interconversion
82	A7FWQ7	CLB_2567	153932267	V-type sodium ATPase, catalytic A subunit	Energy metabolism	ATP-proton motive force interconversion
83	A7FWY1	CLB_2672	153930899	flagellin	Cellular processes	Chemotaxis and motility

84	A7FWY3	CLB_2674	153932791	Flagellar hook-associated protein 2	Cellular processes	Chemotaxis and motility
85	A7FWZ3	CLB_2684	153932820	flagellar motor switch protein FliN	Cellular processes	Chemotaxis and motility
86	A7FX10	CLB_2701	153932227	pyruvate ferredoxin oxidoreductase	Energy metabolism	Electron transport
87	A7FX40	CLB_2737	153931069	peptidase, M29 family	Protein fate	Degradation of proteins, peptides, and glycopeptides
88	A7FX72	CLB_2776	153934192	glycosyl hydrolase, family 18	Energy metabolism	Biosynthesis and degradation of polysaccharides
89	A7FXE1	CLB_2846	153930943	peptidase, M20/M25/M40 family	Protein fate	Degradation of proteins, peptides, and glycopeptides
90	A7FXE4	CLB_2849	153932161	aspartate/ornithine carbamoyltransferase family protein	Unknown function	Enzymes of unknown specificity
91	A7FXL5	CLB_2923	153931346	chaperone protein DnaK	Protein fate	Protein folding and stabilization
92	A7FXU1	CLB_3006	153933712	endoribonuclease L-PSP	Transcription	Degradation of RNA
93	A7FXV9	CLB_3024	153932492	penicillin-binding protein	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan
94	A7FY56	CLB_3136	153932803	oxidoreductase, NAD-binding Rossmann fold family	Unknown function	Enzymes of unknown specificity
95	A7FY57	CLB_3137	153933218	UDP-glucose/GDP-mannose dehydrogenase family	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides
96	A7FY70	CLB_3150	153932809	dehydrogenase, FMN-dependent	Unknown function	Enzymes of unknown specificity
97	A7FY87	CLB_3168	153932923	threonyl-tRNA synthetase	Protein synthesis	tRNA aminoacylation
98	A7FYD2	CLB_3218	153932232	pyrroline-5-carboxylate reductase	Amino acid biosynthesis	Glutamate family
99	A7FYF0	CLB_3236	153934044	thiolase	Fatty acid and phospholipid metabolism	Other
100	A7FYF1	CLB_3237	153933013	3-hydroxybutyryl-CoA dehydrogenase	Energy metabolism	Fermentation
101	A7FYF2	CLB_3238	153932761	3-hydroxybutyryl-CoA dehydratase	Energy metabolism	Fermentation
102	A7FYG4	CLB_3250	153931758	rubredoxin/rubrerythrin	Energy metabolism	Electron transport
103	A7FYI2	CLB_3268	153933152	ATP-dependent Clp protease, proteolytic subunit ClpP	Protein fate	Degradation of proteins, peptides, and glycopeptides
104	A7FYI5	CLB_3271	153931779	conserved hypothetical protein	Hypothetical proteins	Conserved

105	A7FYM5	CLB_3335	153932322	glucose-6-phosphate isomerase	Energy metabolism	Glycolysis/gluconeogenesis
106	A7FYM9	CLB_3340	153933889	D-lactate dehydrogenase	Energy metabolism	Fermentation
107	A7FYN1	CLB_3342	153933291	electron transfer flavoprotein, alpha subunit/FixB family protein	Energy metabolism	Electron transport
108	A7FYN3	CLB_3344	153930801	acyl-CoA dehydrogenase family protein	Fatty acid and phospholipid metabolism	Degradation
109	A7FYN7	CLB_3348	153932598	E-cinnamoyl-CoA:R-phenyllactate CoA transferase	Energy metabolism	Amino acids and amines
110	A7FYP1	CLB_3352	153932667	inosine-5'-monophosphate dehydrogenase	Purines, pyrimidines, nucleosides, and nucleotides	Purine ribonucleotide biosynthesis
111	A7FYP3	CLB_3354	153934242	chaperonin GroEL	Protein fate	Protein folding and stabilization
112	A7FYP4	CLB_3355	153931242	chaperonin, 10 kDa	Protein fate	Protein folding and stabilization
113	A7FYQ6	CLB_3367	153934223	peptidase family protein	Protein fate	Degradation of proteins, peptides, and glycopeptides
114	A7FYS9	CLB_3391	153933536	desulfoferrodoxin	Cellular processes	Detoxification
115	A7FYT9	CLB_3401	153931826	putative dipeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
116	A7FYX3	CLB_3435	153931398	PHP domain protein	Unknown function	General
117	A7FZ12	CLB_3475	153932904	putative L-iditol 2-dehydrogenase	Energy metabolism	Sugars
118	A7FZ15	CLB_3479	153933424	putative 2-oxoacid:acceptor oxidoreductase, alpha subunit	Energy metabolism	Other
119	A7FZ18	CLB_3482	153932770	butyrate kinase	Energy metabolism	Fermentation
120	A7FZ19	CLB_3483	153931375	phosphate butyryltransferase	Energy metabolism	Fermentation
121	A7FZ25	CLB_3489	153934206	oxidoreductase, pyridine nucleotide-disulphide family	Unknown function	Enzymes of unknown specificity
122	A7FZ26	CLB_3490	153930798	thioredoxin	Energy metabolism	Electron transport
123	A7FZ44	CLB_3508	153932349	DNA-directed RNA polymerase, alpha subunit	Transcription	DNA-dependent RNA polymerase
124	A7FZ51	CLB_3515	153932738	methionine aminopeptidase, type I	Protein fate	Protein modification and repair
125	A7FZ71	CLB_3539	153930982	translation elongation factor Tu	Protein synthesis	Translation factors
126	A7FZ72	CLB_3540	153933458	translation elongation factor G	Protein synthesis	Translation factors

127	A7FZ76	CLB_3544	153933989	DNA-directed RNA polymerase, beta' subunit	Transcription	DNA-dependent RNA polymerase
128	A7FZ77	CLB_3545	153933778	DNA-directed RNA polymerase, beta subunit	Transcription	DNA-dependent RNA polymerase
129	A7FZ78	CLB_3546	153932849	50S ribosomal protein L7/L12	Protein synthesis	Ribosomal proteins: synthesis and modification
130	A7FZA5	CLB_3593	153932316	lysyl-tRNA synthetase	Protein synthesis	tRNA aminoacylation
131	A7FZA9	CLB_3597	153932461	pantothenate kinase, type III	Biosynthesis of cofactors, prosthetic groups, and carriers	Pantothenate and coenzyme A
132	A7FZF5	CLB_3705	153933906	CoA-binding protein	Unknown function	General
133	A7FZF7	CLB_3707	153932960	acyl-ACP thioesterase family protein	Fatty acid and phospholipid metabolism	Biosynthesis
134	A7FZH0	CLB_3720	153930842	single-strand binding protein	DNA metabolism	DNA replication, recombination, and repair
135	A7FZI3	CLB_2137	153931821	CoA-transferase, family III	Energy metabolism	Amino acids and amines

96 h

	Replicate A					Replicate B				
Row number	**Protein identification probability**	**Number of unique peptides**	**Number of unique spectra**	**Number of total spectra**	**% cov**	**Protein identification probability**	**Number of unique peptides**	**Number of unique spectra**	**Number of total spectra**	**% cov**
1	100.0%	9	12	141	27.0%	100.00%	8	11	141	23.8%
2	34.4%	1	1	1	5.0%	99.00%	1	1	2	5.8%
3	100.0%	2	2	6	11.8%	100.00%	4	4	12	23.6%
4	99.7%	2	2	45	6.5%	50.00%	1	1	7	3.2%
5	50.0%	1	1	1	4.6%	50.00%	1	1	1	4.6%
6	97.6%	1	1	1	3.4%	100.00%	4	4	8	11.7%
7	100.0%	7	8	39	24.0%	100.00%	4	4	8	13.9%
8	100.0%	5	7	39	14.4%	100.00%	6	7	24	17.1%
9	99.9%	2	2	3	19.0%	99.60%	1	1	4	9.5%
10	100.0%	4	4	24	23.9%	100.00%	4	4	29	23.9%
11	99.8%	2	2	5	11.8%	99.80%	2	2	10	11.8%
12	100.0%	5	7	42	22.5%	50.00%	1	1	3	2.9%
13	99.8%	2	2	6	5.3%	50.00%	1	1	1	3.3%
14	100.0%	3	3	5	10.9%	100.00%	2	2	5	6.6%
15	50.0%	1	1	2	3.5%	50.00%	1	1	4	3.5%
16	100.0%	4	4	10	7.8%	98.00%	1	1	2	1.7%
17	100.0%	6	6	19	6.9%	99.80%	2	2	2	1.8%

18	100.0%	9	12	248	35.8%	100.00%	8	11	106	31.9%
19	100.0%	13	15	62	40.5%	100.00%	10	11	24	31.2%
20	100.0%	5	6	97	28.6%	100.00%	6	7	240	32.7%
21	100.0%	10	11	416	35.5%	100.00%	11	14	206	38.3%
22	100.0%	3	3	11	10.6%	99.30%	1	1	2	2.8%
23	100.0%	26	28	227	35.5%	100.00%	23	26	239	31.2%
24	100.0%	4	4	14	5.3%	100.00%	5	5	22	6.1%
25	50.0%	1	1	2	3.7%	100.00%	3	3	11	8.4%
26	100.0%	3	3	6	16.4%	100.00%	4	4	12	20.7%
27	100.0%	14	14	108	26.0%	100.00%	11	11	96	21.2%
28	99.8%	2	2	5	14.4%	99.70%	2	3	10	14.4%
29	100.0%	5	8	54	20.5%	100.00%	5	6	24	20.5%
30	100.0%	6	7	27	6.4%	100.00%	9	10	64	8.6%
31	100.0%	25	25	84	21.1%	100.00%	24	24	86	20.3%
32	32.2%	1	1	1	7.1%	99.80%	2	2	9	11.5%
33	100.0%	2	2	5	11.0%	100.00%	4	4	12	23.8%
34	100.0%	4	4	11	9.6%	100.00%	3	3	3	7.7%
35	99.6%	1	1	4	2.2%	100.00%	3	3	4	6.3%
36	99.4%	2	2	2	11.5%	99.60%	2	2	3	11.5%
37	100.0%	6	6	9	18.0%	100.00%	8	8	26	23.8%
38	100.0%	2	2	5	5.5%	99.80%	2	2	4	5.5%
39	100.0%	6	7	31	14.3%	100.00%	8	12	68	18.2%

40	100.0%	5	5	17	30.4%	100.00%	5	5	27	30.4%
41	99.8%	2	2	4	32.4%	99.80%	2	2	6	32.4%
42	47.2%	1	1	1	8.2%	100.00%	3	3	3	21.0%
43	25.7%	1	1	1	2.8%	50.00%	1	1	2	2.8%
44	50.0%	1	1	1	3.8%	50.00%	1	1	1	3.8%
45	99.9%	2	2	5	5.7%	99.30%	1	1	2	3.4%
46	100.0%	4	4	9	6.9%	100.00%	4	4	9	6.9%
47	99.8%	2	3	7	13.0%	99.80%	2	2	22	13.0%
48	98.8%	1	1	2	5.6%	99.80%	2	2	8	11.7%
49	100.0%	13	16	169	38.2%	100.00%	9	9	39	26.5%
50	100.0%	12	12	37	10.9%	100.00%	12	12	22	11.6%
51	100.0%	7	7	33	24.5%	100.00%	6	6	23	20.4%
52	50.0%	1	1	1	2.8%	99.90%	2	3	5	7.2%
53	99.8%	2	2	6	13.1%	100.00%	3	3	9	20.2%
54	100.0%	8	8	12	6.5%	100.00%	7	7	12	5.7%
55	100.0%	4	5	14	26.3%	100.00%	4	5	15	26.3%
56	99.7%	2	2	15	6.3%	99.80%	2	2	13	7.1%
57	100.0%	13	14	245	39.4%	100.00%	11	13	232	34.2%
58	100.0%	4	4	11	15.5%	100.00%	5	6	30	19.2%
59	100.0%	4	4	25	21.8%	100.00%	5	5	45	26.2%
60	100.0%	9	10	174	22.9%	100.00%	6	7	127	15.5%

61	99.2%	1	1	1	3.9%	100.00%	3	3	5	9.9%
62	100.0%	10	11	53	17.3%	100.00%	7	8	29	11.7%
63	50.0%	1	1	4	1.5%	100.00%	4	5	11	7.2%
64	99.7%	2	2	2	9.2%	100.00%	2	2	3	8.9%
65	100.0%	4	4	7	5.7%	100.00%	2	2	2	3.0%
66	99.5%	2	2	4	20.2%	99.90%	2	2	12	20.2%
67	100.0%	11	12	38	21.4%	100.00%	12	14	34	22.1%
68	100.0%	3	3	86	44.7%	99.80%	2	2	134	31.8%
69	50.0%	1	1	6	6.1%	99.60%	2	2	3	6.8%
70	98.8%	1	1	7	2.8%	98.70%	1	1	7	2.8%
71	100.0%	9	11	85	32.8%	100.00%	5	5	14	19.4%
72	100.0%	6	6	39	10.8%	100.00%	5	5	21	9.1%
73	99.8%	2	2	8	18.1%	100.00%	3	3	9	27.6%
74	99.8%	2	2	7	17.7%	100.00%	3	3	18	27.2%
75	100.0%	12	17	783	45.9%	100.00%	13	17	435	45.3%
76	100.0%	5	6	48	21.0%	100.00%	3	3	18	13.4%
77	100.0%	6	6	14	12.5%	100.00%	7	7	42	15.3%
78	99.0%	1	1	8	3.7%	99.90%	2	2	6	7.4%
79	100.0%	7	9	36	19.0%	100.00%	6	7	18	17.7%
80	100.0%	2	2	3	14.0%	100.00%	2	2	7	14.0%
81	99.3%	1	1	1	2.8%	50.00%	1	1	2	2.8%
82	100.0%	6	6	9	11.5%	100.00%	7	8	13	13.3%
83	100.0%	6	7	96	30.2%	100.00%	6	6	22	30.2%

84	100.0%	8	8	36	12.8%	100.00%	6	6	13	9.7%
85	99.7%	2	3	3	6.2%	99.70%	2	3	7	6.2%
86	100.0%	23	30	396	22.7%	100.00%	18	22	112	18.0%
87	50.0%	1	1	4	2.4%	99.80%	2	2	2	5.9%
88	99.9%	2	2	8	2.8%	99.80%	2	2	3	2.8%
89	100.0%	6	6	21	20.2%	100.00%	5	6	32	16.8%
90	100.0%	4	5	25	12.1%	100.00%	3	4	7	8.8%
91	100.0%	9	9	131	18.5%	100.00%	4	4	103	7.7%
92	100.0%	4	5	422	42.1%	100.00%	4	6	826	42.1%
93	99.7%	2	2	4	1.9%	100.00%	5	5	9	5.7%
94	100.0%	4	4	20	13.0%	99.80%	2	2	4	6.5%
95	100.0%	7	8	42	19.4%	50.00%	1	1	1	2.7%
96	99.9%	2	2	5	6.2%	99.80%	2	2	5	6.2%
97	100.0%	5	5	8	9.0%	100.00%	4	4	5	7.6%
98	99.8%	2	2	5	9.7%	100.00%	3	3	11	16.8%
99	100.0%	10	12	93	30.6%	100.00%	12	14	107	38.3%
100	100.0%	4	5	89	15.6%	100.00%	4	5	115	15.6%
101	100.0%	7	8	19	31.2%	100.00%	6	7	26	25.8%
102	100.0%	5	6	104	32.6%	100.00%	5	6	189	32.6%
103	100.0%	6	7	51	34.0%	100.00%	4	4	68	23.7%
104	99.6%	2	2	2	9.3%	99.50%	1	1	2	4.1%

105	100.0%	6	7	43	16.0%	100.00%	7	7	20	18.0%
106	100.0%	3	3	11	10.7%	100.00%	3	3	7	10.7%
107	50.0%	1	1	2	3.5%	99.60%	2	2	2	5.8%
108	100.0%	3	3	10	6.9%	100.00%	6	7	38	17.2%
109	100.0%	8	8	37	22.6%	100.00%	3	3	6	8.3%
110	100.0%	8	8	20	21.5%	100.00%	11	12	69	29.8%
111	100.0%	25	30	1718	64.5%	100.00%	23	28	1297	57.9%
112	100.0%	3	3	19	42.1%	100.00%	3	3	9	42.1%
113	100.0%	9	9	46	9.8%	100.00%	13	13	29	13.6%
114	99.3%	1	1	112	10.5%	99.80%	2	2	204	21.0%
115	100.0%	4	4	9	10.8%	99.90%	2	2	4	5.6%
116	50.0%	1	1	2	4.5%	99.50%	2	2	3	10.3%
117	100.0%	3	3	9	11.5%	100.00%	2	2	3	8.3%
118	100.0%	4	4	9	17.1%	100.00%	3	3	6	13.5%
119	100.0%	8	8	30	25.3%	100.00%	12	13	35	41.6%
120	100.0%	7	9	41	29.7%	100.00%	7	8	41	29.7%
121	99.8%	2	2	5	10.5%	99.80%	2	2	4	10.5%
122	100.0%	3	4	31	39.0%	100.00%	3	4	46	39.0%
123	100.0%	8	8	43	29.5%	100.00%	6	6	42	22.9%
124	100.0%	2	2	7	8.0%	100.00%	3	3	8	13.3%
125	100.0%	1	1	10	4.0%	100.00%	2	2	6	6.5%
126	99.9%	1	1	3	1.7%	99.80%	2	2	4	3.6%

127	100.0%	7	7	30	6.4%	100.00%	6	6	14	5.3%
128	100.0%	8	8	35	7.3%	100.00%	7	7	26	6.5%
129	99.6%	2	2	136	17.9%	99.40%	2	2	53	17.9%
130	100.0%	5	5	8	11.5%	100.00%	3	3	6	5.4%
131	99.9%	2	2	3	8.1%	100.00%	2	2	6	10.1%
132	100.0%	3	3	10	28.2%	100.00%	4	4	7	41.9%
133	100.0%	2	2	4	9.2%	100.00%	5	5	6	22.1%
134	99.5%	2	2	3	15.5%	100.00%	3	3	6	23.6%
135	100.0%	12	12	191	3440.0%	100.00%	10	10	75	2900.0%

24 h

Replicate C

Replicate A

Row number	Protein identification probability	Number of unique peptides	Number of unique spectra	Number of total spectra	% cov	Protein identification probability	Exclusive unique peptide count	Exclusive unique spectrum count	Total spectrum count	Protein coverage (%)
1	1	10	13	124	30.4%	100.0%	10	14	108	28.9%
2	1	3	3	4	15.4%	6.1%	1	1	1	5.0%
3	1	2	2	13	11.1%	100.0%	4	4	10	23.6%
4	0.5	1	1	13	3.2%	100.0%	4	4	24	17.5%
5	0.998	2	2	5	9.1%	98.7%	2	2	3	9.1%
6	1	5	5	20	14.3%	100.0%	3	3	5	10.6%
7	1	5	6	13	17.6%	100.0%	11	14	47	40.5%
8	1	6	8	22	17.1%	100.0%	6	8	41	17.1%
9	0.983	1	1	2	9.5%	96.9%	2	2	3	18.1%
10	1	3	3	6	19.0%	100.0%	4	4	20	23.9%
11	0.998	2	2	3	11.8%	50.0%	1	1	2	6.3%
12	1	5	7	19	22.5%	100.0%	5	6	32	22.5%
13	0.5	1	1	1	3.3%	100.0%	3	3	13	8.3%
14	0.5	1	1	2	3.6%	100.0%	3	3	5	10.9%
15	0.998	2	2	4	7.6%	50.0%	1	1	2	3.5%
16	1	3	3	3	5.1%	99.8%	2	2	3	3.6%
17	0.5	1	1	2	1.0%	100.0%	10	11	26	12.2%

18	1	11	14	256	44.8%	100.0%	8	11	200	31.0%
19	1	12	13	46	37.7%	100.0%	15	19	106	49.0%
20	1	7	9	194	37.9%	100.0%	6	8	58	32.7%
21	1	14	18	315	52.2%	100.0%	13	18	133	43.4%
22	1	3	3	10	11.7%	100.0%	5	5	11	21.1%
23	1	23	26	257	32.1%	100.0%	27	34	256	39.4%
24	1	5	5	17	6.1%	100.0%	6	6	19	7.9%
25	1	4	4	16	10.8%	100.0%	4	4	7	12.1%
26	1	3	4	8	16.0%	100.0%	4	4	9	20.7%
27	1	14	14	94	26.8%	100.0%	14	15	102	25.4%
28	0.5	1	2	9	6.8%	99.9%	3	3	10	24.0%
29	1	4	4	41	15.7%	100.0%	5	10	69	20.5%
30	1	9	11	32	9.2%	100.0%	8	12	41	8.0%
31	1	25	25	63	22.3%	100.0%	29	29	100	24.7%
32	1	3	3	7	18.6%	97.4%	2	2	2	11.5%
33	1	2	2	6	11.0%	100.0%	4	5	9	23.8%
34	0.998	2	2	4	6.0%	100.0%	3	4	19	6.4%
35	1	3	3	5	6.3%	100.0%	4	4	14	8.7%
36	1	2	2	2	11.5%	99.9%	3	3	5	16.8%
37	1	7	8	15	21.0%	100.0%	8	9	25	23.8%
38	0.5	1	1	2	3.5%	99.9%	3	3	4	9.0%
39	1	6	8	41	13.7%	100.0%	6	9	38	13.3%

40	1	5	5	24	30.4%	100.0%	5	5	14	30.4%
41	0.998	2	2	8	32.4%	99.8%	2	2	5	32.4%
42	1	4	4	11	26.2%	99.1%	2	2	5	12.8%
43	0.998	2	2	4	7.2%	50.0%	1	1	4	2.8%
44	1	4	4	8	10.5%	98.6%	2	2	3	6.5%
45	0.988	1	1	2	4.3%	100.0%	4	4	9	10.3%
46	1	3	3	6	5.3%	100.0%	7	7	28	13.3%
47	1	3	3	14	21.0%	99.8%	2	3	9	13.0%
48	0.982	1	1	3	5.6%	99.5%	2	2	4	11.7%
49	1	11	13	92	33.1%	100.0%	12	18	145	35.0%
50	1	14	14	49	12.7%	100.0%	19	20	49	17.5%
51	1	7	8	48	24.5%	100.0%	7	9	39	25.5%
52	0.998	2	3	4	7.2%	50.0%	1	1	2	4.5%
53	1	4	4	9	25.1%	99.7%	2	2	5	13.1%
54	1	5	5	9	4.2%	100.0%	12	12	24	9.6%
55	1	3	4	6	19.3%	100.0%	3	4	11	19.3%
56	1	4	4	15	13.4%	99.0%	2	2	15	6.3%
57	1	15	17	345	46.8%	100.0%	13	20	240	41.1%
58	1	6	7	26	25.8%	100.0%	5	6	24	22.1%
59	1	4	5	45	21.8%	100.0%	6	7	34	29.4%
60	1	7	8	135	17.7%	100.0%	10	12	104	25.2%

61	0.998	2	2	4	6.6%	100.0%	4	4	10	13.6%
62	1	7	8	29	11.7%	100.0%	13	16	85	21.1%
63	1	3	5	10	5.7%	99.8%	2	2	4	3.9%
64	1	3	3	8	13.7%	100.0%	4	4	6	18.8%
65	0.999	2	2	3	2.9%	100.0%	8	8	21	10.0%
66	1	2	2	6	20.2%	92.4%	2	2	3	30.9%
67	1	12	14	40	22.3%	100.0%	12	13	45	21.8%
68	0.998	2	2	69	31.8%	50.0%	1	1	6	11.8%
69	0.998	2	2	6	6.8%	97.4%	2	2	4	9.7%
70	0.998	2	2	2	6.7%	99.8%	2	2	7	5.9%
71	1	9	10	34	35.2%	100.0%	13	18	100	49.6%
72	1	3	3	6	5.1%	100.0%	7	7	54	12.7%
73	0.5	1	1	1	8.6%	100.0%	3	3	12	28.4%
74	1	3	3	12	27.2%	99.8%	2	3	5	17.7%
75	1	13	19	983	48.3%	100.0%	13	21	537	48.3%
76	1	5	6	25	21.0%	100.0%	6	8	114	26.1%
77	1	11	12	143	29.7%	100.0%	8	8	50	16.6%
78	0.98	1	1	2	3.7%	98.7%	2	2	8	6.5%
79	1	8	10	40	21.6%	100.0%	9	12	52	25.4%
80	0.997	2	2	4	14.0%	99.4%	2	2	4	14.0%
81	1	3	3	4	9.1%	99.7%	2	2	8	5.9%
82	1	8	10	23	17.2%	100.0%	5	5	16	8.8%
83	1	5	6	34	26.9%	100.0%	7	8	62	33.8%

84	1	8	8	22	13.3%	100.0%	9	10	34	14.2%
85	0.994	1	1	3	3.6%	98.8%	2	2	4	6.2%
86	1	16	21	91	16.0%	100.0%	25	36	354	25.2%
87	0.999	2	2	4	5.9%	50.0%	1	1	4	2.4%
88	0.998	2	2	11	3.5%	100.0%	4	4	9	6.1%
89	1	7	9	39	24.0%	100.0%	9	11	44	32.3%
90	1	5	5	23	14.3%	100.0%	4	7	48	12.1%
91	1	4	4	30	8.3%	100.0%	12	15	140	24.4%
92	1	4	5	371	42.1%	100.0%	4	7	233	42.1%
93	1	3	3	3	3.6%	100.0%	4	4	8	4.3%
94	0.999	2	2	4	6.5%	100.0%	7	7	19	26.0%
95	1	5	6	14	13.7%	100.0%	5	6	35	12.8%
96	1	2	2	4	9.2%	50.0%	1	1	3	3.3%
97	1	4	4	10	6.6%	100.0%	5	5	14	8.2%
98	1	5	5	18	25.0%	50.0%	1	1	2	4.1%
99	1	14	18	156	51.8%	100.0%	13	16	112	49.7%
100	1	5	6	209	22.1%	100.0%	9	11	43	42.4%
101	1	4	4	9	17.7%	100.0%	7	8	21	32.7%
102	1	6	8	192	40.9%	100.0%	4	5	57	26.5%
103	1	6	8	117	31.4%	100.0%	7	9	41	37.6%
104	0.5	1	1	1	5.2%	50.0%	1	1	2	4.1%

105	1	6	7	43	16.0%	100.0%	8	9	55	22.0%
106	0.5	1	1	2	3.3%	100.0%	4	4	11	14.0%
107	0.5	1	1	2	3.5%	99.6%	2	2	6	5.8%
108	1	11	13	55	38.2%	100.0%	3	3	23	6.9%
109	1	6	6	12	18.0%	100.0%	8	9	43	22.3%
110	1	7	7	31	19.4%	100.0%	12	13	39	29.5%
111	1	23	27	1506	61.0%	100.0%	22	32	1151	54.2%
112	0.995	1	1	3	10.5%	100.0%	3	3	18	31.6%
113	1	12	12	25	12.6%	100.0%	16	16	50	16.6%
114	0.997	2	2	94	21.0%	50.0%	1	1	49	10.5%
115	0.991	1	1	5	1.9%	100.0%	8	8	29	23.5%
116	0.5	1	2	4	4.5%	99.8%	2	3	5	11.1%
117	0.998	2	2	3	6.0%	100.0%	4	4	14	12.6%
118	0.998	2	3	16	9.6%	100.0%	4	5	16	17.1%
119	1	13	14	62	44.1%	100.0%	11	11	44	36.0%
120	1	7	9	59	29.7%	100.0%	8	10	61	34.0%
121	1	3	3	6	15.3%	99.8%	2	2	6	9.4%
122	1	4	5	19	49.5%	99.8%	2	2	10	23.8%
123	1	9	9	54	32.4%	100.0%	12	14	53	44.8%
124	1	2	2	4	8.0%	100.0%	4	4	6	17.3%
125	1	3	4	13	8.8%	100.0%	7	8	23	23.2%
126	0.999	2	2	3	3.6%	100.0%	6	6	17	9.7%

127	1	7	7	19	7.1%	100.0%	14	14	46	14.2%
128	1	8	8	22	7.2%	100.0%	13	13	48	12.0%
129	0.5	1	1	3	11.4%	99.2%	2	3	21	17.9%
130	1	2	3	4	4.6%	100.0%	7	7	20	15.9%
131	1	2	2	2	10.1%	99.8%	2	2	3	7.4%
132	0.996	2	2	3	16.9%	100.0%	3	4	7	28.2%
133	1	4	4	8	19.3%	100.0%	4	4	9	19.3%
134	0.987	1	2	3	8.1%	50.0%	1	1	1	7.4%
135	1	11	11	69	3120.0%	100.0%	15	19	185	4210.0%

Replicate C

Row number	Exclusive unique peptide count	Exclusive unique spectrum count	Total spectrum count	Protein coverage (%)	Protein identification probability	Exclusive unique peptide count	Exclusive unique spectrum count	Total spectrum count	Protein coverage (%)
1	10	14	83	28.9%	100.0%	10	16	92	30.9%
2	2	2	6	10.8%	99.8%	2	2	8	10.8%
3	4	4	25	23.6%	100.0%	4	4	27	23.6%
4	1	1	6	3.2%	99.8%	2	2	13	5.8%
5	2	2	4	9.1%	99.9%	3	3	6	12.2%
6	6	6	27	18.0%	100.0%	6	6	35	18.0%
7	8	9	21	30.7%	100.0%	9	10	18	31.5%
8	6	7	22	17.1%	100.0%	6	8	17	17.1%
9	2	2	5	18.1%	100.0%	3	3	5	27.6%
10	4	4	26	23.9%	100.0%	3	3	7	19.0%
11	2	2	3	9.7%	50.0%	1	1	2	6.3%
12	1	1	3	2.9%	100.0%	3	4	8	12.1%
13	1	1	4	3.3%	50.0%	1	1	2	3.3%
14	4	4	12	13.8%	100.0%	3	3	5	10.9%
15	1	1	6	3.5%	99.8%	2	2	4	7.6%
16	4	4	7	6.7%	37.5%	1	1	2	1.6%
17	2	2	4	1.8%	50.0%	1	1	3	1.0%

18	8	11	136	31.0%	100.0%	10	13	233	40.6%
19	14	15	40	44.7%	100.0%	16	19	96	49.5%
20	7	8	134	37.9%	100.0%	7	9	144	37.9%
21	10	14	68	29.9%	100.0%	15	22	186	51.7%
22	4	4	7	14.7%	100.0%	4	4	9	15.3%
23	23	31	179	30.6%	100.0%	21	27	244	27.5%
24	6	6	36	7.9%	100.0%	5	5	24	6.1%
25	5	6	16	14.2%	100.0%	5	6	17	14.2%
26	4	4	23	20.7%	100.0%	5	6	18	25.8%
27	16	17	85	28.4%	100.0%	16	17	102	28.3%
28	3	4	14	24.0%	97.6%	2	2	11	16.4%
29	5	7	57	20.5%	100.0%	5	7	56	20.5%
30	8	9	36	6.8%	100.0%	7	8	26	6.1%
31	27	27	73	21.8%	100.0%	22	22	50	18.1%
32	3	3	8	18.6%	100.0%	3	4	10	18.6%
33	4	5	18	23.8%	99.9%	3	3	8	16.6%
34	2	3	6	4.7%	100.0%	3	4	13	6.4%
35	4	4	8	8.7%	100.0%	6	6	14	12.8%
36	2	2	3	11.5%	99.8%	2	2	7	11.5%
37	6	6	22	18.0%	100.0%	7	8	19	20.8%
38	1	1	2	2.1%	50.0%	1	1	1	3.5%
39	8	11	69	17.4%	100.0%	8	9	74	16.8%

40	3	3	23	15.7%	100.0%	3	3	13	15.7%
41	3	3	8	54.9%	99.8%	2	2	7	32.4%
42	2	2	8	12.8%	100.0%	3	3	15	17.4%
43	3	3	4	8.8%	99.8%	2	2	3	5.7%
44	1	1	1	3.8%	100.0%	3	3	9	8.3%
45	2	2	5	4.8%	100.0%	4	4	9	10.3%
46	7	8	22	13.3%	100.0%	6	8	13	11.2%
47	3	3	23	19.1%	100.0%	3	4	19	19.1%
48	3	3	8	16.8%	100.0%	3	3	8	16.8%
49	11	14	49	32.6%	100.0%	11	13	72	32.6%
50	15	15	21	13.9%	100.0%	17	18	49	15.2%
51	8	10	36	26.5%	100.0%	11	13	62	37.6%
52	2	3	6	7.2%	99.2%	2	2	5	7.2%
53	4	5	18	25.1%	100.0%	4	6	14	25.1%
54	9	9	16	7.1%	100.0%	8	8	13	6.3%
55	4	4	11	26.3%	100.0%	3	3	8	19.3%
56	3	4	16	10.1%	99.8%	2	3	8	6.3%
57	14	20	226	43.9%	100.0%	15	23	405	45.8%
58	5	5	25	19.2%	100.0%	6	7	38	25.8%
59	7	7	21	32.9%	100.0%	7	9	30	32.9%
60	10	13	123	25.2%	100.0%	10	13	110	25.2%

61	2	2	4	6.0%	100.0%	3	3	6	9.9%
62	11	13	63	17.9%	100.0%	11	12	51	17.9%
63	3	4	7	5.7%	100.0%	5	5	13	9.4%
64	3	3	9	13.7%	100.0%	4	5	8	18.8%
65	7	7	15	8.7%	100.0%	5	5	11	6.9%
66	1	1	2	20.2%	100.0%	3	3	6	40.4%
67	16	19	48	29.8%	100.0%	15	16	68	28.1%
68	1	1	10	11.8%	99.8%	2	2	9	29.4%
69	2	2	4	7.4%	40.5%	1	1	7	6.8%
70	3	3	11	9.8%	100.0%	3	3	5	9.8%
71	7	8	26	25.1%	100.0%	10	15	54	37.9%
72	6	6	16	10.4%	99.9%	3	3	9	5.0%
73	4	4	12	37.9%	99.6%	2	2	4	18.1%
74	2	3	9	17.7%	99.8%	2	3	9	17.7%
75	14	21	429	48.6%	100.0%	15	25	665	53.8%
76	5	5	19	21.3%	100.0%	6	9	84	26.1%
77	10	11	49	20.4%	100.0%	9	10	62	18.2%
78	3	5	9	10.2%	100.0%	5	6	9	17.9%
79	11	13	34	30.8%	100.0%	10	12	51	27.5%
80	2	2	5	14.0%	99.8%	2	2	4	14.0%
81	1	1	2	2.8%	100.0%	6	6	9	17.8%
82	7	8	26	11.7%	100.0%	7	9	34	11.7%
83	6	6	25	30.2%	100.0%	8	9	32	38.9%

84	8	8	14	12.9%	100.0%	9	9	31	14.2%
85	2	2	7	6.2%	99.3%	2	2	5	6.2%
86	17	22	119	16.0%	100.0%	13	15	99	11.8%
87	2	2	4	5.4%	99.8%	2	2	3	5.9%
88	4	4	6	5.5%	100.0%	8	8	20	11.6%
89	8	12	42	28.1%	100.0%	10	15	50	35.6%
90	4	7	41	12.1%	100.0%	4	7	50	12.1%
91	6	7	82	11.9%	100.0%	6	7	45	12.2%
92	4	7	490	42.1%	100.0%	4	6	283	42.1%
93	4	4	12	4.5%	100.0%	4	4	5	4.3%
94	4	4	7	14.9%	100.0%	6	6	10	21.7%
95	2	2	2	4.6%	100.0%	6	6	15	15.8%
96	3	3	4	9.2%	50.0%	1	1	4	3.3%
97	6	7	9	9.9%	100.0%	5	6	11	8.2%
98	2	3	13	8.2%	100.0%	3	4	8	13.1%
99	16	21	151	57.1%	100.0%	18	22	166	69.1%
100	8	12	72	36.2%	100.0%	10	15	125	45.3%
101	7	8	21	33.1%	100.0%	5	5	8	23.1%
102	6	9	130	39.2%	100.0%	6	8	98	39.2%
103	4	7	78	23.2%	100.0%	7	10	113	39.2%
104	3	3	7	14.1%	99.2%	2	2	4	9.3%

105	9	11	32	24.0%	100.0%	9	11	39	24.0%
106	2	2	9	6.5%	98.7%	2	2	3	6.5%
107	2	2	7	5.8%	99.8%	2	2	5	5.8%
108	8	8	60	22.3%	100.0%	11	13	59	35.0%
109	6	6	10	16.7%	100.0%	10	10	32	28.2%
110	9	12	85	23.6%	100.0%	10	12	44	26.9%
111	20	28	801	47.5%	100.0%	20	29	1149	48.1%
112	4	4	22	52.6%	50.0%	1	1	4	10.5%
113	15	15	31	15.3%	100.0%	10	10	19	9.9%
114	2	3	119	21.0%	98.4%	2	2	97	21.0%
115	3	3	6	7.6%	100.0%	6	6	17	16.4%
116	2	2	6	11.1%	99.8%	2	2	7	11.1%
117	3	3	5	9.5%	100.0%	3	3	8	9.5%
118	1	2	9	3.9%	99.7%	2	3	17	9.6%
119	12	14	29	39.6%	100.0%	14	17	57	50.6%
120	9	11	48	37.6%	100.0%	9	12	70	38.0%
121	3	3	6	12.5%	100.0%	6	7	15	30.3%
122	4	5	28	49.5%	100.0%	4	4	14	49.5%
123	6	6	41	21.0%	100.0%	11	13	54	40.0%
124	4	4	9	17.3%	99.6%	3	3	4	14.9%
125	5	5	8	17.1%	100.0%	7	8	22	23.2%
126	4	4	8	6.2%	100.0%	3	3	5	4.9%

127	7	7	19	6.5%	100.0%	9	9	17	8.4%
128	9	10	28	8.0%	100.0%	10	10	30	8.8%
129	2	3	24	17.9%	50.0%	1	1	6	11.4%
130	3	3	7	6.2%	100.0%	6	6	15	12.3%
131	4	4	9	16.3%	100.0%	3	3	8	13.6%
132	3	4	12	28.2%	99.6%	2	2	3	17.7%
133	5	5	16	22.5%	100.0%	4	5	11	16.9%
134	2	3	6	15.5%	15.8%	1	1	2	8.1%
135	14	17	105	3840.0%	100.0%	15	18	102	4230.0%

Appendix table 4: Proteins identified in the culture supernatant of *C. botulinum* A1 19397 at 96 h only, Uniprot accession nu

Uniprot accession	Locus	GI number	Common name	Main functional role	Functional sub role
A7FZ55	CLB_3523	153933040	30S ribosomal protein S8	Protein synthesis	Ribosomal proteins: synthesis and modification
A7FZ05	CLB_3468	153933429	putative transaldolase	Energy metabolism	Pentose phosphate pathway
A7FYW5	CLB_3427	153933908	oxidoreductase, aldo/keto reductase family	Unknown function	Enzymes of unknown specificity
A7FVX7	CLB_2266	153934009	stage V sporulation protein S	Cellular processes	Sporulation and germination
A7FVE6	CLB_2043	153931453	bacterial microcompartments family protein	Cellular processes	Detoxification
A7FUN1	CLB_1762	153931445	conserved hypothetical protein	Hypothetical proteins	Conserved
A7FU43	CLB_1551	153931027	glyoxalase family protein	Unknown function	Enzymes of unknown specificity
A7FR77	CLB_0503	153932318	lemA family protein	Unknown function	General
A7FPT1	CLB_0491	153932733	hydroxyethylthiazole kinase	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine

mber, common name, gene locus, main role and sub role

Replicate A					Replicate B				
Protein identification probability	Number of unique peptides	Number of unique spectra	Number of total spectra	% coverage	Protein identification probability	Number of unique peptides	Number of unique spectra	Number of total spectra	% coverage
50.0%	1	1	5	9.8	99.8%	2	2	5	19.7
99.6%	2	2	2	11.6	99.8%	2	2	4	11.1
50.0%	1	1	2	4.5	43.2%	1	1	1	4.5
99.8%	2	2	4	38.4	99.8%	2	2	8	38.4
50.0%	1	1	2	14.6	99.9%	2	3	6	25.0
97.0%	1	1	1	11.7	99.5%	2	2	2	20.4
50.0%	1	1	3	12.4	100.0%	3	3	18	32.2
99.8%	2	2	3	13.2	99.7%	2	2	3	13.2
99.8%	2	2	4	9.9	99.8%	2	2	6	9.9

Replicate C

Protein identification probability	Number of unique peptides	Number of unique spectra	Number of total spectra	% coverage
50.0%	1	1	2	9.8
100.0%	3	3	3	17.3
99.7%	2	2	2	8.3
98.7%	1	1	1	20.9
50.0%	1	2	2	14.6
50.0%	1	1	1	8.7
99.8%	2	2	6	22.3
50.0%	1	1	2	6.3
99.8%	2	2	4	9.9

Appendix Table 5: Proteins identified in the culture supernatant of *C. botulinum* B NCTC 7273

Row number	Uniprot accession	Locus	GI number	Common name	Main role	Sub role
1	B1ID40	CLD_0720	170756148	arginine deiminase	Energy metabolism	Amino acids and amines
2	B1ID49	CLD_0711	170755122	methionyl-tRNA synthetase	Protein synthesis	tRNA aminoacylation
3	B1ID73	CLD_0687	170756964	isoleucyl-tRNA synthetase	Protein synthesis	tRNA aminoacylation
4	B1IDB5	CLD_0549	170756620	glyceraldehyde-3-phosphate dehydrogenase, type I	Energy metabolism	Glycolysis/gluconeogenesis
5	B1IDB6	CLD_0548	170754551	phosphoglycerate kinase	Energy metabolism	Glycolysis/gluconeogenesis
6	B1IDB7	CLD_0547	170755354	triosephosphate isomerase	Energy metabolism	Glycolysis/gluconeogenesis
7	B1IDB8	CLD_0546	170756209	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	Energy metabolism	Glycolysis/gluconeogenesis
8	B1IDB9	CLD_0545	170757014	enolase	Energy metabolism	Glycolysis/gluconeogenesis
9	B1IDD5	CLD_0373	170755070	putative cell surface protein	Cell envelope	Other
10	B1IDI4	CLD_0341	170756703	clpB protein	Protein fate	Protein folding and stabilization
11	B1IEF5	CLD_0304	170757296	peptidase T	Protein fate	Degradation of proteins, peptides, and glycopeptides
12	B1IEH9	CLD_0280	170754553	lemA family protein	Unknown function	General
13	B1IF05	CLD_0407	170756566	aldehyde-alcohol dehydrogenase	Energy metabolism	Fermentation
14	B1IF54	CLD_0207	170754585	Xaa-pro aminopeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
15	B1IFC4	CLD_1235	170757106	acyl-CoA dehydrogenase family protein	Fatty acid and phospholipid metabolism	Degradation

16	B1IFC8	CLD_1231	170756655	E-cinnamoyl-CoA:R-phenyllactate CoA transferase	Energy metabolism	Amino acids and amines
17	B1IFD2	CLD_1227	170757052	inosine-5'-monophosphate dehydrogenase	Purines, pyrimidines, nucleosides, and nucleotides	Purine ribonucleotide biosynthesis
18	B1IFD4	CLD_1225	170756380	chaperonin GroEL	Protein fate	Protein folding and stabilization
19	B1IFD5	CLD_1224	170754876	chaperonin, 10 kDa	Protein fate	Protein folding and stabilization
20	B1IFE7	CLD_1212	170755674	peptidase family protein	Protein fate	Degradation of proteins, peptides, and glycopeptides
21	B1IFH0	CLD_1188	170756958	desulfoferrodoxin	Cellular processes	Detoxification
22	B1IFH6	CLD_1182	170754655	zinc metalloprotease, aminopeptidase I family	Protein fate	Degradation of proteins, peptides, and glycopeptides
23	B1IFI1	CLD_1177	170755947	putative dipeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
24	B1IFN3	CLD_2582	170757156	myosin-cross-reactive antigen family protein	Cell envelope	Surface structures
25	B1IFS9	CLD_2536	170756226	enhancing factor	Unknown function	General
26	B1IG10	CLD_1084	170756406	putative 2-oxoacid:acceptor oxidoreductase, beta subunit	Energy metabolism	Other
27	B1IG11	CLD_1083	170756108	putative 2-oxoacid:acceptor oxidoreductase, alpha subunit	Energy metabolism	Other
28	B1IG12	CLD_1082	170755301	putative 2-oxoacid:acceptor oxidoreductase, delta subunit	Energy metabolism	Other
29	B1IG14	CLD_1080	170757474	butyrate kinase	Energy metabolism	Fermentation
30	B1IG15	CLD_1079	170755600	phosphate butyryltransferase	Energy metabolism	Fermentation
31	B1IG22	CLD_1072	170754359	oxidoreductase, pyridine nucleotide-disulphide family	Unknown function	Enzymes of unknown specificity

32	B1IG23	CLD_1071	170754340	thioredoxin	Energy metabolism	Electron transport
33	B1IG68	CLD_2484	170756053	putative PTS system, L-Ascorbate family, IIB component	Signal transduction	PTS
34	B1IGC5	CLD_1053	170756437	DNA-directed RNA polymerase, alpha subunit	Transcription	DNA-dependent RNA polymerase
35	B1IGD7	CLD_1041	170755311	30S ribosomal protein S5	Protein synthesis	Ribosomal proteins: synthesis and modification
36	B1IGE0	CLD_1038	170757480	30S ribosomal protein S8	Protein synthesis	Ribosomal proteins: synthesis and modification
37	B1IGF6	CLD_1008	170754754	translation elongation factor Tu	Protein synthesis	Translation factors
38	B1IGG2	CLD_1016	170757275	DNA-directed RNA polymerase, beta subunit	Transcription	DNA-dependent RNA polymerase
39	B1IGJ2	CLD_0972	170757532	lysyl-tRNA synthetase	Protein synthesis	tRNA aminoacylation
40	B1IH03	CLD_0937	170756872	stage V sporulation protein G	Cellular processes	Sporulation and germination
41	B1IHB0	CLD_3660	170757546	conserved hypothetical protein	Hypothetical proteins	Conserved
42	B1IHF2	CLD_2343	170755724	aminoacyl-histidine dipeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
43	B1IHI4	CLD_2311	170755141	metallopeptidase, family M24	Protein fate	Degradation of proteins, peptides, and glycopeptides
44	B1IHZ7	CLD_3552	170757216	pyridoxal-phosphate-dependent aminotransferase family	Amino acid biosynthesis	Aspartate family
45	B1II28	CLD_2243	170755446	phosphocarrier protein HPr	Signal transduction	PTS
46	B1II33	CLD_2238	170756258	stage V sporulation protein S	Cellular processes	Sporulation and germination
47	B1II65	CLD_2206	170757774	translation elongation factor Ts	Protein synthesis	Translation factors
48	B1IIF5	CLD_3478	170757004	putative glycosyl hydrolase	Energy metabolism	Biosynthesis and degradation of polysaccharides

49	B1IIH5	CLD_2175	170756865	D-proline reductase, PrdA proprotein	Energy metabolism	Amino acids and amines
50	B1III5	CLD_2164	170755390	proline racemase	Energy metabolism	Amino acids and amines
51	B1IIK8	CLD_2140	170754397	conserved hypothetical protein	Hypothetical proteins	Conserved
52	B1IIM1	CLD_2127	170754628	peptide deformylase	Protein fate	Protein modification and repair
53	B1IIU4	CLD_3422	170756660	putative thiosulfate sulfurtransferase	Central intermediary metabolism	Sulfur metabolism
54	B1IIW8	CLD_3398	170756609	putative methenyltetrahydrofolate cyclohydrolase	Central intermediary metabolism	One-carbon metabolism
55	B1IJB8	CLD_3332	170756597	deoxyuridine 5'-triphosphate nucleotidohydrolase	Purines, pyrimidines, nucleosides, and nucleotides	2'-Deoxyribonucleotide metabolism
56	B1IJD9	CLD_3311	170755165	glycine reductase, subunits ABC	Energy metabolism	Amino acids and amines
57	B1IJH7	CLD_1991	170755126	selenium metabolism protein YedF	Unknown function	General
58	B1IJJ6	CLD_1972	170757461	ornithine carbamoyltransferase	Amino acid biosynthesis	Glutamate family
59	B1IJJ7	CLD_1971	170755451	carbamate kinase	Energy metabolism	Amino acids and amines
60	B1IJK8	CLD_1960	170755811	peptidase, M20/M25/M40 family	Protein fate	Degradation of proteins, peptides, and glycopeptides
61	B1IJM2	CLD_1946	170756857	metallo-beta-lactamase family protein/flavodoxin	Energy metabolism	Electron transport
62	B1IJM5	CLD_1943	170756373	peptide deformylase	Protein fate	Protein modification and repair
63	B1IJM7	CLD_1941	170754730	V-type sodium ATPase, B subunit	Energy metabolism	ATP-proton motive force interconversion
64	B1IJM8	CLD_1940	170757651	V-type sodium ATPase, catalytic A subunit	Energy metabolism	ATP-proton motive force interconversion
65	B1IJR0	CLD_3275	170755356	transaldolase	Energy metabolism	Pentose phosphate pathway
66	B1IJZ5	CLD_1907	170754574	flagellar hook capping protein	Cellular processes	Chemotaxis and motility
67	B1IK59	CLD_1843	170755201	flagellin	Cellular processes	Chemotaxis and motility

Row number	Exclusive unique peptide count	Exclusive unique spectrum count	Total spectrum count	Protein coverage (%)
1	1	1	1	2.2%
2	2	2	4	4.0%
3	1	1	2	1.0%
4	10	13	121	40.0%
5	13	17	126	42.7%
6	8	10	248	41.1%
7	3	3	8	8.8%
8	14	19	682	51.0%
9	42	52	2292	39:8%
10	4	4	19	5.3%
11	7	8	35	19.4%
12	2	2	3	13.2%
13	22	28	240	29.5%
14	3	4	14	9.2%
15	9	10	76	28.9%

16	4	4	12	10.9%
17	7	9	43	18.8%
18	23	38	932	60.3%
19	2	2	14	31.6%
20	3	3	6	3.4%
21	2	3	463	21.0%
22	4	4	9	12.0%
23	2	3	7	6.3%
24	6	8	20	10.6%
25	2	2	4	2.6%
26	4	5	32	18.9%
27	2	2	5	9.6%
28	2	2	11	27.5%
29	12	14	36	41.0%
30	6	7	29	25.4%
31	5	5	17	21.3%

32	4	6	125	49.5%
33	2	2	5	20.2%
34	6	6	25	21.9%
35	1	1	1	5.5%
36	1	1	4	9.8%
37	6	6	22	18.6%
38	2	2	6	1.8%
39	5	5	7	10.7%
40	2	2	13	24.2%
41	4	4	32	15.1%
42	8	8	16	20.7%
43	13	15	40	24.0%
44	2	2	6	6.6%
45	2	3	201	31.8%
46	2	2	3	38.4%
47	1	1	3	2.9%
48	4	5	19	9.4%

49	2	2	7	2.7%
50	4	5	8	15.2%
51	1	1	3	8.6%
52	3	3	13	27.2%
53	6	6	18	13.5%
54	2	2	9	11.0%
55	2	2	3	11.5%
56	5	5	16	15.2%
57	2	2	13	12.4%
58	14	22	607	50.5%
59	2	2	3	9.2%
60	14	14	304	34.4%
61	8	10	39	21.6%
62	2	2	3	14.0%
63	9	10	30	29.7%
64	13	16	40	27.7%
65	6	6	21	35.0%
66	2	2	7	8.4%
67	7	8	288	37.3%

68	7	7	31	10.7%
69	1	1	1	16.4%
70	3	3	4	10.7%
71	2	2	19	3.2%
72	10	12	34	10.7%
73	2	2	7	5.9%
74	1	1	2	2.3%
75	4	4	7	8.8%
76	4	6	10	9.3%
77	3	3	4	15.1%
78	3	3	4	21.7%
79	1	1	1	4.0%
80	2	2	7	13.0%
81	8	8	14	11.8%
82	6	6	27	19.2%
83	4	5	60	31.8%
84	2	2	5	8.4%

85	12	14	150	36.3%
86	1	1	4	13.2%
87	3	3	6	5.8%
88	17	18	52	15.5%
89	7	9	66	26.0%
90	4	6	1249	42.1%
91	1	1	1	0.6%
92	2	2	3	13.5%
93	15	22	973	43.2%
94	1	1	1	4.8%
95	8	8	17	22.1%
96	5	6	26	22.9%
97	3	3	7	8.4%
98	3	3	6	15.5%
99	1	1	1	9.2%
100	10	11	406	25.4%
101	4	4	13	7.4%
102	2	2	6	7.8%
103	12	15	144	38.5%

104	3	4	13	12.3%
105	3	3	5	13.1%
106	5	5	11	8.3%
107	3	3	12	5.5%
108	7	9	309	49.2%
109	5	6	55	28.9%
110	9	12	41	24.4%
111	2	2	5	6.6%
112	3	3	6	10.6%
113	23	25	84	20.2%
114	15	19	68	15.1%
115	8	14	231	29.6%
116	2	3	80	13.0%
117	18	18	227	32.4%
118	2	2	3	250.0%